Universidade de Santiago de Compostela
Departamento de Física de Partículas e Instituto Galego de Física de Altas Enerxías

Bayesian analysis of the mass composition of Ultra-High Energy Cosmic Rays using X_{max} data recorded at the Pierre Auger Observatory

Guillermo Torralba Elipe

September 2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Universidade de Santiago de Compostela

Doctoral Thesis

Bayesian analysis of the mass composition of Ultra-High Energy Cosmic Rays using X_{max} data recorded at the Pierre Auger Observatory

Author:
Guillermo Torralba Elipe

Supervisor:
Dr. Enrique Zas Arregui

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Physics in the

Departamento de Física de Partículas e Instituto Galego de Fsica de Altas Energxías

September 2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Universidade de Santiago de Compostela
Departamento de Física de Partículas e Instituto Galego de Física de Altas Enerxías

Enrique Zas Arregui, catedrático de la Universidade de Santiago de Compostela,
CERTIFICA:

que la memoria titulada **Bayesian analysis of the mass composition of Ultra-High Energy Cosmic Rays using $X_{\max}$ data recorded at the Pierre Auger Observatory** es el trabajo realizado, bajo mi supervisión, por Guillermo Torralba Elipe en el Departamento de Física de Partículas e Instituto Galego de Física de Altas Enerxás y que constituye el trabajo de tesis que presenta para obtar al grado de Doctor en Física.

Fdo: Enrique Zas Arregui

September 2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Declaration of Authorship

I, Guillermo Torralba Elipe, declare that this thesis titled, **Bayesian analysis of the mass composition of Ultra-High Energy Cosmic Rays using $X_{\max}$ data recorded at the Pierre Auger Observatory** and the work presented in it are my own. I confirm that when I use the work of others it is referenced.

Yo, Guillermo Torralba Elipe, declaro que esta tesis titulada **Bayesian analysis of the mass composition of Ultra-High Energy Cosmic Rays using $X_{\max}$ data recorded at the Pierre Auger Observatory** y el trabajo que en ella se presenta es mío. Confirmo que cuando uso el trabajo de otros está claramente referenciado.

Signed: Guillermo Torralba Elipe

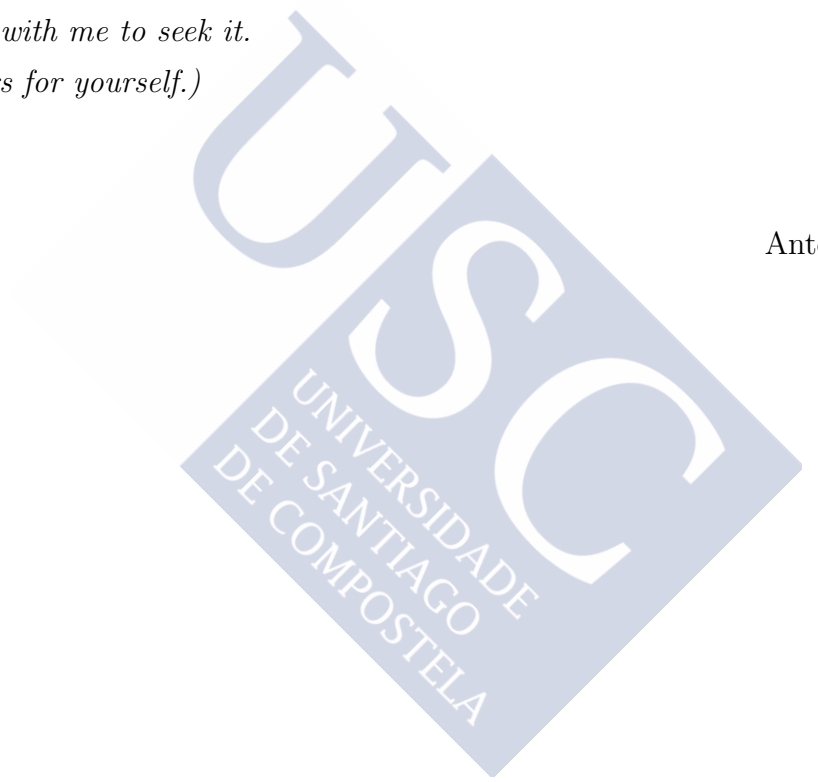
September 2017

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

"¿Tu verdad? no, la verdad;
y ven conmigo a buscarla.
La tuya guárdatela."

(Your truth? no,
Truth;
and come with me to seek it.
Keep yours for yourself.)

Antonio Machado



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Acknowledgements

I want to thank to all of my partners of the Pierre Auger Collaboration and members that are and were in the University of Santiago de Compostela such as A. Yushkov, S. Riggi, G. Rodrǵuez, J. Alvarez-Muñiz and G. Parente.

I am specially grateful to E. Zas for his endless corrections and comments; and R. A. Vázquez and I. Valiño for their continuous assistance and fruitful discussions.

Of course, this work could not be done without the unconditional support of my family.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Preface

The goal of this work is the determination of the composition of the Ultra-High-Energy Cosmic Rays arriving to the Earth using the $X_{\max}$ distributions. In CHAPTER 1 we briefly introduce the cosmic rays, observables aimed for the composition analysis, the Pierre Auger Observatory and its recent results. The reason for this brief description is the large amount of literature and Ph.D thesis already written about these topics. We make a summary of the Bayesian statistical inference in CHAPTER 2 in order to clarify the terminology used in this work. In CHAPTER 3 we explore different statistical estimators for the composition analysis. CHAPTER 4 and CHAPTER 5 are reserved respectively for the presentation of the data used for the composition analysis and the analysis itself. In CHAPTER 6 we use the information of the composition obtained in CHAPTER 5 to establish a preliminary proton flux measured at the Earth. Finally, in CHAPTER 7 we present the conclusions of the work.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

To my family

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 1

Introduction

In this chapter we comment general features of the cosmic rays and we focus more on the Ultra-High-Energy Cosmic Rays. We briefly describe the candidates for sources of these particles and the main interactions that they can experience until their reach the Earth.

We also describe minimally the main observables for the composition analysis, the Pierre Auger Observatory and the recent results presented by the Pierre Auger Collaboration.

1.1 Cosmic rays and particle astrophysics: general remarks

Cosmic rays are charged particles and nuclei of extraterrestrial origin that are continuously reaching the Earth, discovered by Victor Hess in 1912 [1]. More than a century after the discovery they continue being object of study and of highest priority in modern astrophysics. Cosmic-ray energies cover a wide range from several MeV up to around 10^{20} eV, being the highest-energy cosmic ray ever detected the event with an energy of $3.2 \cdot 10^{20}$ eV observed in 1991 by the Fly's Eye Collaboration [2]. The cosmic rays at the end of the energy spectrum, those with energies above 10^{18} eV, are called Ultra-High-Energy Cosmic Rays (UHECRs) and constitute the most-energetic particles in the universe. Understanding where and how these particles are accelerated up to these extremely large energies is the main motivation of the

study of UHECR, opening a unique window to the unknown, most violent phenomena in the universe. These particles also provide unique insights into particle physics interactions at energies much above those achieved in artificial accelerators. These facts make the field of UHECR into an exciting field of research at the intersection of elementary particle physics and astrophysics. Indeed, the field of elementary particle physics owes its origin to discoveries made in course of cosmic-ray research.

The answer to the questions *how* and *where* the UHECRs are produced is tied to the own nature of these particles. As it is summarised through this chapter, the knowledge of the mass and charge of such particles is essential to understand and interpret the observations. Since the flux of UHECRs at Earth is extremely low, direct measurements are not possible, and data comes from large ground-based experiments that measure extensive showers of secondary particles initiated in the interaction of the UHECR in the atmosphere. In such indirect measurements, the information about composition is limited by our theoretical understanding of the hadronic interaction properties at such high energies. These facts explain why in spite of a wealth of measurements already done by large experiments around the world, the origin and nature of UHECRs is still largely unknown. For a historical review of observational studies of cosmic rays see e.g. [3].

1.1.1 The energy spectrum of UHECRs

One of the most important observational results about cosmic rays is the measurement of the all-particle energy spectrum, which carries combined information about the sources of cosmic rays and about galactic and/or intergalactic media in which cosmic rays propagate.

The cosmic-ray energy spectrum falls by 25 orders of magnitude over 11 decades of energy, being well described by a power law with an index which is a function of energy. Being measured by a number of cosmic-ray experiments, three prominent spectral features have been clearly observed as shown in FIGURE 1.1: a steepening at $\sim 3{\times}10^{15}$ eV known as the *knee* a flattening, the widely recognised *ankle*, at $5{\times}10^{18}$ eV; and the abrupt suppression of the flux at energies beyond $\sim 5{\times}10^{19}$ eV. Other (less pronounced) two spectral features have been also reported: a flattening at $\sim 10^{16}$ eV called the *low-energy ankle*; and a steepening at $\sim 10^{17}$ eV known as the *second knee*;

Assuming that the low-energy cosmic rays are of galactic origin, the *knee* is commonly believed to be caused by the maximum energy of acceleration available at the Galactic sources, and by the maximum energies of the magnetic confinement for protons in the Galaxy. The *second knee* reported by KASCADE-Grande experiment at an energy of 8×10^{16} eV about 26 times higher than the knee, is consistent with the idea that the knee structures are rigidity-dependent cut-offs and therefore their positions are proportional to the charge of the nuclei. The galactic cosmic-ray sources would be the so-called "Pevatrons", reaching a maximum energy for particles with charge Ze of $E/Z \approx 3\times10^{15}$ eV. Alternative scenarios that interpret knee-like structures as an unexpected change of the hadronic interaction cross sections at shower level are strongly disfavoured by the experimental results.

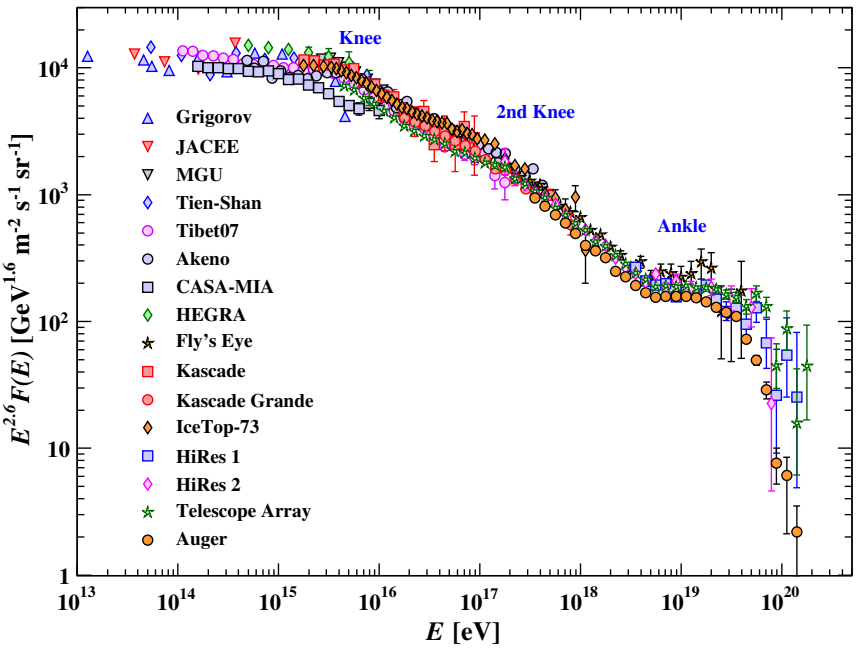


FIGURE 1.1: The all-particle spectrum from air showers measurements [4]. The data are scaled with $E^{2.6}$ to better present the spectral features.

Above 10^{18} eV, the regime of the UHECRs, it is widely believed that the cosmic rays are of extragalactic origin, since the galactic magnetic fields can not confine them in the Galaxy (see SECTION 1.1.2), supported by the level of nearly isotropy observed in the arrival directions of UHECRs. The *ankle* feature can be interpreted as the result of an extragalactic flux (the harder component) beginning to dominate over

the galactic flux (the softer component) or as the imprint of dominant extragalactic protons suffering e^+e^- proton pair-production processes in the cosmic microwave background (CMB). In the latter scenario, known as the *dip* [5], the transition from galactic to extragalactic cosmic rays must begin at a lower energy, possibly around the *second knee* and be completed before the *dip*, and it is characterised by a sharp change of mass composition from galactic iron to extragalactic protons [6].

Regarding the *low-energy ankle*, the hardening in the flux reported by Yakutsk [7] and IceCube [8] collaborations at $\sim 10^{16}$ eV, this could be interpreted as the imprint of a second component of cosmic rays of galactic origin, in addition to a standard galactic component expected to be produced at the supernova remnants (SNR), that experience a charge-dependent of cut-offs [9].

The only firm prediction ever made concerning the shape of the UHECR spectrum was made in 1966 by Greisen [10], and independently by Zatsepin and Kuz'min[11]. They predicted a suppression of the cosmic-ray flux around $5{\times}10^{19}$ eV due to energy losses by photon-pion production from the interaction of the CRs with the low energy CMB photons. This feature is known as the GZK cut-off. In the case of a mixed composition, the photo-disintegration of heavy nuclei have a similar effect. Another alternatives of the observed flux suppression are related to the maximum energy of acceleration attainable by the sources of UHECRs.

Although the energy spectrum carries powerful pieces of information about the origin and propagation of UHECRs, its measurement by itself, despite the high level of precision reached by current experiments, does not allow one to conclude unambiguously about the origin of the spectral features and thereby about the origin of UHECRs. The correct interpretation of the spectral features requires the additional key information from the measurement of the primary composition of UHECRs.

1.1.2 Propagation of UHECRs

While the UHECRs travel from their sources to the Earth, they are affected by two type of processes (for a review see e.g. [12, 13]): the interactions with cosmic background; and the interaction with cosmic magnetic fields. The interactions with the CMB, the infrared, optical and ultraviolet (IR-UV) photons change the energy and mass of these particles while the interactions with the magnetic fields only deflect their trajectories (and therefore affect the time that the particles take to travel from

their sources to the Earth). Both leave a variety of imprints on the observables of UHECRs such as the energy spectrum and primary composition observed on Earth.

Interactions with cosmic backgrounds

At the UHECR energies, the extragalactic radiation fields relevant for cosmic ray interactions are the CMB at the highest energies and the IR-UV photons (also known as extragalactic background light, EBL) at slightly lower energies.

In the case of ultra-high energy (UHE) protons the main interaction processes are photo-pion production and the e^+e^- pair-production on the CMB:

$$p\gamma \longrightarrow \Delta^+ \longrightarrow N + n\pi \tag{1.1}$$

$$p\gamma \longrightarrow p\, e^+\, e^-. \tag{1.2}$$

In Equation 1.1, N is a nucleon and n is the number of pions produced (well above the resonance energy for Δ^+ production, multi-π production dominates). The energy thresholds of these processes for a photon of energy ϵ are: $\sim 1.2{\times}10^{20}$ eV (ϵ_{CMB}/ϵ) for photo-pion production, and $\sim 0.8{\times}10^{18}$ eV (ϵ_{CMB}/ϵ) for pair production, with $\epsilon_{CMB} \sim 6{\times}10^{-4}$ eV as the mean energy of a CMB photon and assuming head-on collisions. As a result of both interactions the proton spectrum is distorted. Pair-production (Equation 1.2) produces the *dip* feature at energies $1{\times}10^{18} - 4{\times}10^{19}$ eV in an extragalactic proton spectrum. If UHE protons originate at cosmological distances, photo-pion production (Equation 1.1) produces the GZK cut-off which induces a sharp drop of the flux above an energy $E_{GZK} \simeq 5{\times}10^{19}$ eV [10, 11]. At about $5{\times}10^{19} - 10^{20}$ eV a bump in the proton spectrum is expected to be caused by the higher energy protons that have lost energy and pile-up where the photo-pion production cross section starts decreasing, after which the proton spectrum declines steeply [14]. The GZK effect implies that almost all protons arriving on Earth with energies above the threshold must come from sources closer than ~ 100 Mpc, which is known as the GZK horizon. Note that the shape, size and position of the expected features depend on the shape of the injection spectrum, cosmic-ray luminosity, cosmological evolution and distribution of the cosmic rays sources in the Universe.

In the case of UHE nuclei with atomic mass number A the energy threshold for photo-production is higher by a factor A, but photo-disintegration on both EBL and

CMB dominates at lower energies (mainly through the giant dipole resonance) :

$$A\gamma \longrightarrow (A-n)\,nN \tag{1.3}$$

with n being the number of nucleons emitted. This process changes the nuclei species giving rise to the production of secondary nuclei or/and nucleons. The resulting nuclear fragments may be unstable and decay and speeding up the energy loss of the whole nucleus. Nuclei also experience pair production that decreases their Lorenz factor without affecting their composition:

$$A\gamma \longrightarrow A\,e^{+}\,e^{-}. \tag{1.4}$$

For these process only CMB field is relevant. For a detailed study of the propagation of UHE nuclei through CMB and EBL and its impact on the observed spectra see e.g. [15, 16].

One remarkable effect of the nuclei propagation is that cosmic rays with mass number $A < 20$ can not travel farther than few tens of Mpc without disintegrating, while protons and iron nuclei may reach Earth from sources at distances up to about 100 Mpc [17]. This implies that heavy nuclei could be found in abundance at the suppression energies only if the composition at the sources was basically dominated by iron-group (or heavier) nuclei.

In the energy range of interest, UHECR propagate over cosmological distances losing energy adiabatically due to the expansion of the Universe with a typical energy loss length of the order of 4 Gpc.

Therefore, one can conclude the actual shape and position of the GZK feature in the all-particle spectrum depend on the characteristics of the sources and on their local spatial distribution, and also on the cosmic-ray composition. Regarding the ankle feature (see Section 1.1.1), like the GZK cut-off, it could be directly linked to interaction of protons with CMB photons [5] if UHECRs were dominated by protons. Pair-production suffered by nuclei is not expected to imprint any feature in the energy spectrum. Note that as we will see later in this thesis the composition inferred by the Pierre Auger Collaboration and that obtained in this work indicates that also intermediate elements are present at the energy at which ankle feature appears.

Interactions with the magnetic fields

Charged particles are subject to the influence of magnetic fields in the source environment, in the intergalactic medium, and in the Galaxy. As it will be shown in the following, there are different regimes of propagation depending on the strength of the magnetic field, the composition and the energy of the UHECRs.

A charged particle with Lorentz factor Γ, mass m and charge Ze moving perpendicular to a magnetic field B describes a helicoidal trajectory whose angular frequency (ω_L) and cyclotron radius or Larmor radius (r_L) are given by EQUATION 1.5 and EQUATION 1.6, respectively:

$$\omega_L = \frac{ZeB}{\Gamma mc}, \tag{1.5}$$

$$r_L = \frac{pc}{ZeB} \simeq \frac{E}{ZeB}. \tag{1.6}$$

In EQUATION 1.6 the particle is assumed to travel with a velocity close to the speed of light. The ratio between the momentum and the electric charge of the particle is called *rigidity* R:

$$R = \frac{pc}{Ze} \simeq \frac{E}{Ze}. \tag{1.7}$$

The rigidity is the quantity governing the motion of the particle inside the magnetic field B. A proton with energy E will describe the same trajectory that an iron with energy $E \times 26$.

Our galaxy can be described roughly as a spheroidal halo of radius ~ 30 kpc and a disk of radius ~ 15 kpc and thickness of ~ 300 pc ([18]) filled with a magnetic field of $\sim 4\mu$G parallel to the spiral arms. The Larmor radius can be rewrite as

$$r_L(Mpc) = \left(\frac{1.08}{Z}\right)\left(\frac{\text{nG}}{B}\right)\left(\frac{E}{\text{EeV}}\right). \tag{1.8}$$

In TABLE 1.1 some Larmor radius for different proton energies are shown.

E	R_L
0.1	0.03
1	0.3
10	3
100	30

TABLE 1.1: Rounded Larmor radius (R_L) in kpc for a proton with four different energies (in EeV) travelling in the Galaxy.

Comparing these results with the thickness of the Galaxy one can observe that a proton originates inside the Galaxy escapes if its energy is larger than 1 EeV.

As it has been commented, the magnetic fields change the directions of the charged particles. Assuming a regular magnetic field the deflection angle of a particle with charge Z crossing perpendicularly after a distance d is

$$\theta \simeq 0.5^{\circ} Z \left(\frac{100 \text{ EeV}}{E}\right)\left(\frac{d}{\text{kpc}}\right)\left(\frac{B}{\mu\text{G}}\right). \tag{1.9}$$

The halo is believed to have a regular and a turbulent component. The latter is thought to be coherent, *i.e.*, well defined, within regions of scale l_c but the fields of the different coherent regions are randomly oriented.

Similarly extragalactic space is thought to have random fields of order nG and with a coherent length of order 1 Mpc. In such scenario deviations from sources a distance d, not too far away from us, correspond to relatively small angles:

$$\langle\theta\rangle \simeq 0.8^{\circ} Z \left(\frac{100 \text{ EeV}}{E}\right)\left(\frac{B}{\text{nG}}\right)\sqrt{\frac{d}{10 \text{ Mpc}}}\sqrt{\frac{l_c}{\text{Mpc}}}. \tag{1.10}$$

Small angular deflections are attractive because in such case by detecting the cosmic ray direction we can roughly infer the source position.

It must however be stressed that the experimental evidence related to the magnetic fields only serves to poorly constrain them. This is particularly true for the extragalactic fields. The angular deviations are moreover linearly dependent on Z (i.e. on primary composition) so all together highly uncertain.

The point is that it is plausible that deviations are small and hence that we can infer source positions from anisotropies in the arrival directions. Indeed some such anisotropies have been already reported (see below Section 1.4). In such case the knowledge of composition and the source positions would provide most valuable information to infer the properties of intervening magnetic fields much more accurately. Besides the relevance of understanding the sources of UHECRs this would also be an enormous step forward in establishing the nature and extent of the galactic and extragalactic magnetic fields

1.1.3 Origin of UHECRs

The most popular candidates for the sources of the bulk of the galactic cosmic rays (GCR) are Supernova Remnants (SNR). Cosmic rays are believed to be accelerated by diffusive shock waves (known as the first-order Fermi mechanism) of expanding SNRs in the Galaxy reaching energies up to $\sim$ few $Z\times10^{15}$ eV [19], which corresponds to the region where the *knee* feature in the spectrum occurs. Diffusive shock-wave acceleration model applied to SNRs explains quite naturally the rigidity-dependence associated to the knee-structures as previously discussed in SECTION 1.1.1, but it excludes not only UHECRs, but also the higher GCRs. For most other Galactic sources, the energy reached is estimated to be too small to explain the UHECRs energies with the exception of pulsars and magnetars which could produce a relatively hard spectra [20].

The scale for such maximum acceleration was set up by Hillas [21]. It stems from the basic requirement for a potential source that the magnetic field strength B is strong enough to trap the cosmic-ray particle of charge Ze within the accelerating region over which it extends, of linear size L. The maximum acceleration energy is estimated requiring the Larmor radius of the particle (given by EQUATION 1.8) to be less than the accelerating region ($L \geq 2\,r_L/\beta$), obtaining the following expression:

$$E_{max} \simeq Ze\,\beta \left(\frac{B}{\mu\mathrm{G}}\right)\left(\frac{L}{\mathrm{kpc}}\right)[10^{18}\mathrm{eV}]. \tag{1.11}$$

FIGURE 1.2 is a version of the original figure of Hillas showing few objects that satisfy the conditions B and L needed to achieve energies exceeding 10^{20} eV. Note that the criterion illustrated by this diagram is a necessary, but not sufficient condition and would imply maximal efficiency. Additional constraints further limit the maximum energies than can be achieved and the potential sources of UHECR acceleration. A powerful test for identifying the sources of UHECRs would be the observation of gamma-ray and/or neutrino fluxes from the candidates. The sources of CRs should also emit high-energy photons (neutrinos) resulting from decays of secondary neutral (charged) pions produced when the primary protons/nuclei interact with the surrounding gas and/or radiation fields.

Among the most promising classes of extragalactic astrophysical accelerators of UHECRs are active galactic nuclei (AGN) and Startburst galaxies. Both classes of

sources match the energy production rate in UHECRs estimated to be of order 10^{44}erg Mpc^{-3}yr-1 [22].

Starburst galaxies are those undergoing a massive star formation episode. They feature strong emissions associated to interstellar extinction, massive stars as gamma-ray bursts (GRB), hypernovae and magnetars. From such an active region a galactic-scale superwind is driven by the collective effect of supernovae and particularly massive star winds. Startburst galaxies would have the power to efficiently accelerate heavy nuclei up to energies even beyond the GZK energy limit [23].

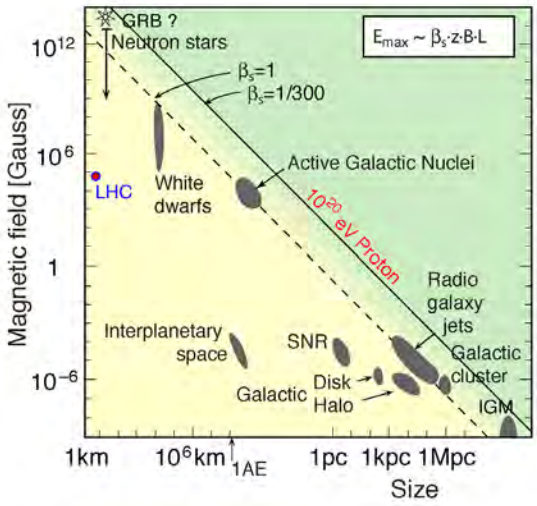


FIGURE 1.2: Hillas diagram that shows the size and magnetic field strength of astrophysical objects that are candidate sources for UHECRs. The solid (dashed) line corresponds to the condition to accelerate protons (iron) at 10^{20} eV.

The discussion on the possible sources of UHECRs and the propagation effects that those particles suffer in their travel from the sources to the Earth brings to the forefront the prominent role that the knowledge of the mass composition of UHECR plays, and how important it is to unveil their origin. The measurement of the energy-dependent composition of UHECR is one of the main goals of this thesis. The results in this work (see CHAPTER 5) allow us to infer the UHE proton spectrum from the measured all-particle spectrum (see CHAPTER 6) and to provide us with hints on which scenarios of UHECR origin are more plausible.

1.2 Cosmic Ray detection

1.2.1 Extensive Air Showers

When a (primary) nucleus with ultra-high energy interacts with the molecules of the atmosphere produces secondary particles with less energy than the primary. These particles interact again with atmospheric nuclei producing more particles and the cycle is repeated until the secondaries reach an energy at which they do not decay or lose energy and the multiplication process stops. The whole process is called Extensive Air Shower (EAS). An illustration is given in FIGURE 1.3.

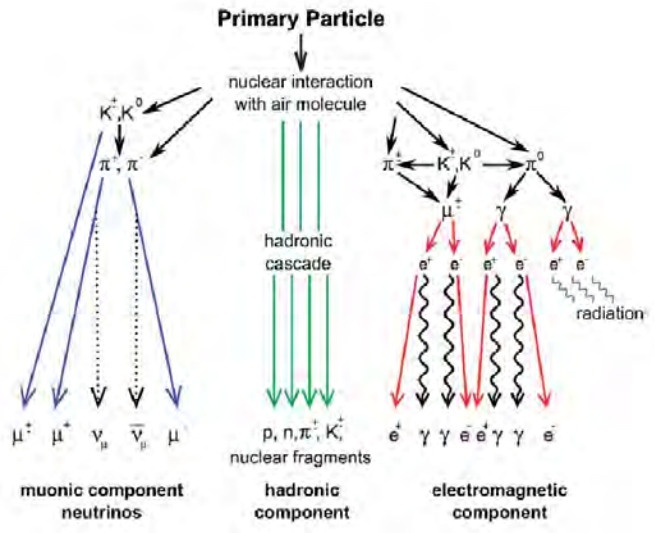


FIGURE 1.3: Illustration of an EAS initiated by a nucleus.

The products of the first interaction (the collision of the primary particle with the atmosphere) are multiple particles, mainly charged and neutral mesons (π and K) together with few heavier hadrons, that will follow equivalent interactions with less energy. Since the K-mesons decay in π-mesons we can approximate the first products to only π-mesons. In such way, we can model the EAS as the combination of three components: the muonic cascade (composed by muons and neutrinos), the electromagnetic cascade (composed by electrons and photons) and the hadronic cascade which keeps feeding the other two components.

A simple analytic model that attempts to describe the EAS was built by Heitler [24], but this only accounted for the development of the electromagnetic cascade. Subsequent approaches to incorporate the hadronic component and so to extend Heitler's model to the case of EAS were developed [25]. These models predict the main characteristics of the EAS by describing the shower as a branching process. The primary particle interacts in the atmosphere and is split into different secondary particles as illustrated in FIGURE 1.4.

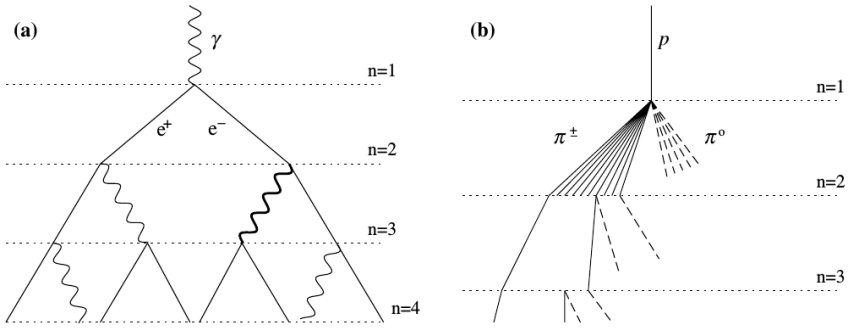


FIGURE 1.4: Illustration of the EAS branching process. Electromagnetic shower generated by a photon in panel (a) and a shower generated by a proton in panel (b). The proton produces secondary neutral and charged pions. Neutral pions decay into two photons producing the shower (a) and charged pions further interact and induce the shower (b).

The branching process stops when the energy of secondary particles reach a critical energy. For electromagnetic showers only $e^{\pm}$ and photons are generated while in the proton-induced EAS neutral and charged pions are produced with the same multiplicity ($N_{\pi^0} = N_{\pi^+} = N_{\pi^-}$) and the primary proton energy is also equally distributed. Neutral pions decay into photons ($\pi^0 \rightarrow \gamma\gamma$) producing an electromagnetic shower while the charged pions interact in the atmosphere producing other neutral and charged pions. The electromagnetic sub-shower induced by π^0 stops when the energy of particles drops to $\simeq 85$ MeV in air. The sub-shower produced by $\pi^{\pm}$ ceases when the energy of the pions drops below a critical energy E^{π}_{dec} at which it begins to become more likely that charged pions decay rather than interact. At this energy all charged pions are assumed to decay into muons and neutrinos ($\pi^+ \rightarrow \mu^+\nu_\mu$; $\pi^- \rightarrow \mu^-\bar{\nu}_\mu$).

Particularly relevant for the composition analysis (and thus for this work) is the behaviour of the the depth at which the shower reaches the maximum number of

particles, $X_{\max}$. For a proton-initiated shower, $X_{\max}$ can be expressed as follows:

$$X^p_{\max} \simeq \lambda^p_I + X_0 \ln\left(\frac{E_0}{2NE^\pi_{dec}}\right). \tag{1.12}$$

Here, λ^p_I is the depth of the first proton interaction, X_0 is the interaction length of the secondary particles (that is assumed to be constant), E_0 is the primary proton energy and N is the total number of pions produced in the first interaction ($N = N_{\pi^0} + N_{\pi^+} + N_{\pi^-}$).

Note that apart from $X_{\max}$ there are other EAS measurable properties that can be used as mass-tracers and therefore predictions for such observables are also relevant for the composition analysis. As said above, muons are produced in the decay of $\pi^\pm$ when their energies drops below E^π_{dec}. Assuming that all charged pions decay, the number of muons produced in a proton-induced shower is given by:

$$N^p_\mu = \left(\frac{E_0}{E^\pi_{dec}}\right)^\beta, \tag{1.13}$$

where $\beta = \ln(N_{ch})/\ln(N) \simeq 0.9$.

The simplest way to describe a shower initiated by a cosmic-ray nucleus with mass number A and energy E_0 is to apply this branching model together with the *superposition model*. In the latter, the shower induced by a nucleus is assumed to be equal to the superposition of A separate proton showers, each of them with energy E_0/A. Under this assumption the number of muons produced at the end of the shower is:

$$N^A_\mu = A\left(\frac{E_0/A}{E^\pi_{dec}}\right)^\beta = A^{1-\beta}N^p_\mu, \tag{1.14}$$

and the depth of shower maximum of the electromagnetic component is:

$$X^A_{\max} \simeq \lambda^A_I + X_0 \ln\left(\frac{E_0}{2ANE^\pi_{dec}}\right). \tag{1.15}$$

From Equation 1.14 and Equation 1.15 one can express the resulting shower properties in terms of A and of the corresponding quantities of a proton shower as:

$$\begin{cases} N^A_\mu = N^p_\mu A^{1-\beta} \\ X^A_{\max} \simeq X^p_{\max} - \left[(\lambda^p_I - \lambda^A_I) + X_0 \ln A\right] \end{cases}. \tag{1.16}$$

From these expressions one can notice the following. On the one hand the number of muons in a nucleus-induced shower is larger than the number of muons in a proton-induced shower, at the same total primary energy. On the other hand since $\lambda_I^p > \lambda_I^A$ the $X_{\max}$ of a nucleus-induced shower is smaller than that induced by a proton.

Taking only geometrical assumptions one can express $\lambda_I^A \simeq \lambda_I^p / A^{2/3}$. Therefore the standard deviation of the $X_{\max}$ distribution produced by a nucleus of $A > 1$ is also smaller than the standard deviation of that produced by a proton. The properties of the $X_{\max}$ distribution that show its capability as mass tracer, confirmed with detailed simulations, can be summarised as:

$$\begin{cases} \langle X_{\max}^A \rangle < \langle X_{\max}^p \rangle \\ \sigma(X_{\max}^A) < \sigma(X_{\max}^p) \end{cases} . \tag{1.17}$$

Both number of muons reaching the ground and $X_{\max}$ are good measurable variables to study the nature of the primary particle that generates the shower, nevertheless $X_{\max}$ is more sensitive to the composition. Thus, $X_{\max}$ is the observable that we will use in the following in this work for the composition analysis.

1.2.2 Detection techniques

For a good review of the different detection techniques the reader can see [26], [27] and [28]. At ultra-high energies the flux of cosmic rays at Earth is so low that direct measurements are not possible, and it is necessary to measure the properties of the UHECR, such as energy, direction, and mass, by indirect measurements of the extensive air shower induced in the atmosphere. The most standard methods to study the UHECRs consist in *ground arrays* and *fluorescence telescopes*.

Ground arrays

It was the first technique to measure the EAS (in fact, they were discovered thanks to this method [29]), it consist in the distribution of particle detectors over a large area to detect the secondary particles produced in the EAS that reach the ground. Ground arrays include arrays of scintillators, muon detectors, water-Cherenkov detectors, etc. When these particles reach the ground their densities and times are registered by the detectors. The arrival time is used to estimate the direction of the shower while the distribution of the signals or particle densities is used to estimate the shower size

which is proportional to the energy. Usually the relation between signal and energy is obtained from detailed simulations and it depends on the primary composition.

In addition this technique is subject to simulation uncertainties because they require making assumptions on the particle interactions in regions not explored with particle accelerators. The advantage is that it has a duty cycle of 100% what makes it the method that can register most events reaching large exposures and reducing the statistical uncertainties.

Fluorescence telescopes

The molecules of nitrogen of the atmosphere are ionised with the passage of the shower particles and when they deexcite, they emit fluorescence light isotropically ([30]). The photons of the light correspond to the energies of the transition levels of the nitrogen molecules. The light emitted is proportional to the energy deposition and the advantage of this technique is that it registers the longitudinal profile of the shower. The integral of the total emitted light gives a calorimetric measurement of the electromagnetic energy released by the shower in the atmosphere. The total primary energy is then derived by taking into account corrections due to the attenuation of light in the atmosphere due to the absorption and scattering of the photons and adding an estimate of the energy carried into the ground by high-energy muons and neutrinos which is not deposited int the atmosphere and does not contribute to the light emission. The timing of the detection of the photons in the pixels of the cameras together with the directional information of the pixels themselves gives the shower direction. The disadvantage of this technique is the need of a clean atmosphere and a moonless night resulting in a poor duty cycle of order of 10%.

1.3 The Pierre Auger Observatory

The Pierre Auger Observatory [31] was conceived as the largest and most precise detector to measure the spectrum and arrival directions of UHECRs with unprecedented precision in an attempt it to establish their origin. It was designed to take the advantage of the two techniques mentioned above (hybrid detection) by combining a large surface detector (SD) array and a fluorescence detector (FD). It exploits the large aperture of the SD, operating continuously, as well as the measurement of the shower development in the atmosphere obtained with the FD (duty cycle of 13%). The hybrid design allows one to measure (indirectly) the parameters explained in

Section 1.2.1 (with minimal use of simulations) helping us to distinguish between the primary particles that initiated the showers.

The Pierre Auger Observatory is located in the province of Mendoza, Argentina (centered at 69°20 W, 35°20 S) at a mean altitude of 1400 m above the sea level[1] and covers an area of 3000 km^2 which makes it the largest man-made detector of history. The area is filled with the SD, composed of a baseline array of 1600 water-Cherenkov detectors (WCD) separated by 1500 m arranged by parallelogram unitary cells. The surface store is overlooked by 24 fluorescence telescopes, the FD, located at 4 perimeter buildings. In addition, there is a smaller nested array of 61 additional WCD stations spaced by 750 m (covering an area of 24 km^2) and three additional fluorescence detectors (HEAT). The Observatory is depicted in Figure 1.5.

The Pierre Auger Observatory has been collecting high-quality data since 2004, running with its full configuration since 2008. It has already led to key measurements that have dramatically advanced our understanding of UHECRs. A review of selected results is made in Section 1.4, with emphasis given to the measurement of the energy spectrum and mass composition due to their relevance to the work presented in this thesis.

In this chapter we only give a brief description of the Surface and Fluorescence detectors. A full description of the design and performance of all detector systems, associated infrastructure as well as the ongoing upgrade of the Auger Observatory can be found in [31, 32].

[1] 1400 m above the sea level corresponds to an atmospheric overburden of $\sim 875\ g\,cm^{-2}$.

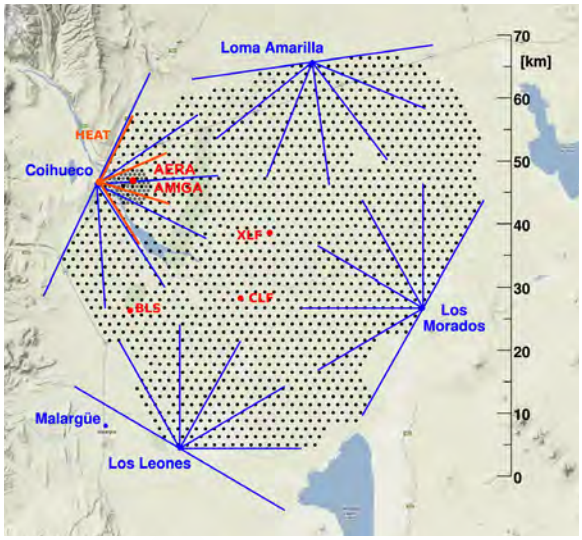


FIGURE 1.5: The Pierre Auger Observatory layout. Each dot corresponds to one of the surface detector stations. The four fluorescence detectors sites are shown, each with blue lines marking the 30° field of view of its six telescopes. The orange lines correspond to the HEAT telescopes. Also shown are other facilities of the Observatory (see [31] for more details).

1.3.1 Overview of the surface detector

Each WCD station consists of a cylindrical tank of 3.6 m diameter and 1.55 m height, enclosing a sealed liner filled with 12,000 L of ultra-pure water. The liner is coated with a reflective surface on the inside. Above the tank, there are three 9 inches photo-multiplier tubes (PMTs) looking into the water to collect the Cherenkov light from the passage of relativistic charged particles through the water. The WCD is also sensitive to high-energy photons that convert to e^+e^- pairs in the water volume. Each PMT provides two signals that are digitised in time slots of 25 ns by a pair of 10 bit 40 MHz semi-flash Analog to Digital Converters (ADC). The digital data are clocked into a programmable logic device, which is used to monitor the ADC outputs for local trigger patterns. Each station is a stand-alone system. There is a solar power system that supplies power to the PMTs and electronics package. A schematic view of a WCD station is shown in FIGURE 1.6.

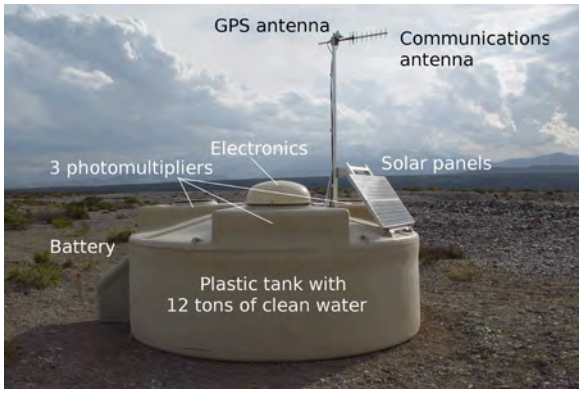


FIGURE 1.6: A schematic view of a surface detector station in the field, with the main components labelled.

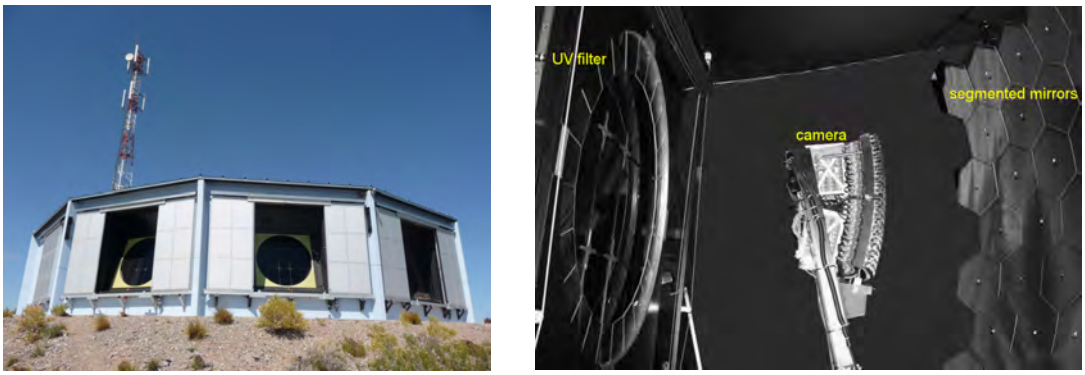
The signal measured by each WCD station is normalised to a common calibration unit both to cancel out the dependence on detector parameters and to calibrate against the detector simulations. This unit is called the *vertical equivalent muon* (VEM) and corresponds to the signal induced by a vertical muon traversing the detector in a vertical trajectory. These muons give a characteristic signal peak when all the signals are recorded in a given station and can thus be easily identified. All recorded signals are converted into VEM units prior to data analysis.

A complex SD trigger system is implemented to select high-quality extensive air showers from the background of atmospheric muons. This is hierarchical system with two low level triggers implemented by the local front-end electronics and a third level trigger formed at the central data acquisition system (CDAS) based on spatial and temporal correlation of the lower triggers (the central trigger). Additional higher trigger levels can be implemented offline to select high-quality physical events.

1.3.2 Overview of the fluorescence detector

The 24 telescopes of the FD are located in 4 sites at the perimeter of the SD array: Los Leones, Los Morados, Loma Amarilla and Coihueco. At each site there are six independent telescopes in a clean climate-controlled building, as the one shown in panel (A) of FIGURE 1.7. Each telescope has a field of view (FoV) of $30°\times30°$ in azimuth and elevation, with a minimum elevation of $1.5°$ above the horizon. The three additional HEAT telescopes with an elevated FoV (from $30°$ to $58°$ in elevation) are about 180 m in front of the FD site at Coihueco. The HEAT telescopes allow to

observe showers induced by cosmic rays with energies below the second knee up to the ankle.



(A) (B)

FIGURE 1.7: (A) Photo of the FD building at Los Leones during the day (with open shutters due to maintenance). (B) A schematic view of a fluorescence telescope, with the main components labelled.

A schematic view of fluorescence detector telescope is shown in panel (B) of FIGURE 1.6. The fluorescence light enters through a circular diaphragm covered with a UV-filter glass window, is focused by a 13 m^2 spherical segmented mirror and detected by a camera formed by 440 PMTs, whose single FoV is $\sim 1.5°$

The calibration of the FD telescopes in terms of photons at aperture per ADC count in the PMTs is achieved by approximately yearly absolute calibrations. The molecular properties of the atmosphere at the time of data taking are obtained with local weather stations. The aerosol content of the atmosphere and the amount of clouds are continuously monitored.

As the PMT data are processed, they are passed through a flexible trigger system implemented in firmware and software. The accepted events are sent to a central readout computer that builds an event from the coincident data in all the telescopes at a given site and generates a hybrid trigger for the surface detector. The use of the timing measurement at at least a single station of the SD is enough to allow a precise determination of the shower axis in space, key step towards a high-quality measurement of the longitudinal profile.

1.4 Recent progress in the field done by the Pierre Auger Observatory

In the last decade, the results of the Pierre Auger Observatory have dramatically advanced our understanding of UHECRs. However they have also lead to a number of puzzling observations that indicate a complex astrophysical scenario. For instance, the most recent physics results can be found in [33–41]. Here we focus on the most relevant outcomes related to the scope of this thesis.

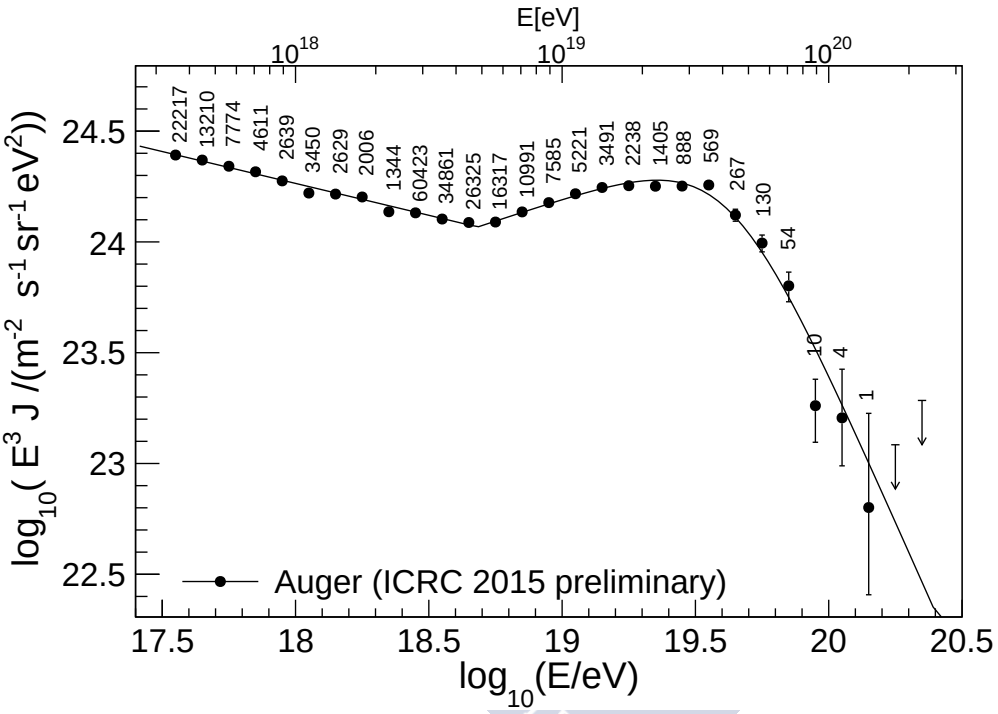


FIGURE 1.8: The all-particle energy spectrum of cosmic rays (multiplied by E^3) as measured by the Auger Observatory. Only statistical uncertainties are shown. The systematic uncertainty on the energy is 14%.

The all-particle cosmic-ray flux above $3{\times}10^{17}$ eV has been measured using the data collected for more than 10 years [42]. The differential energy spectrum, depicted in FIGURE 1.8, is obtained by combining four independent data sets. The *ankle* is found to be at an energy of $(4.82{\pm}0.07{\pm}0.8(\text{sys})){\times}10^{18}$ eV, and the flux suppression above $(42.1{\pm}1.7{\pm}7.6(\text{sys})){\times}10^{18}$ eV is unquestionably established with a significance of more than 20σ.

In spite of having measured these features with unprecedented precision, the origin of UHECRs is still unknown. Several scenarios can successfully explain them, as

discussed in SECTION 1.1.1. The Pierre Auger Collaboration has addressed this challenge in parallel through measurements of the mass composition of the observed data and through studies of the distribution of the arrival directions of the primaries over the sky.

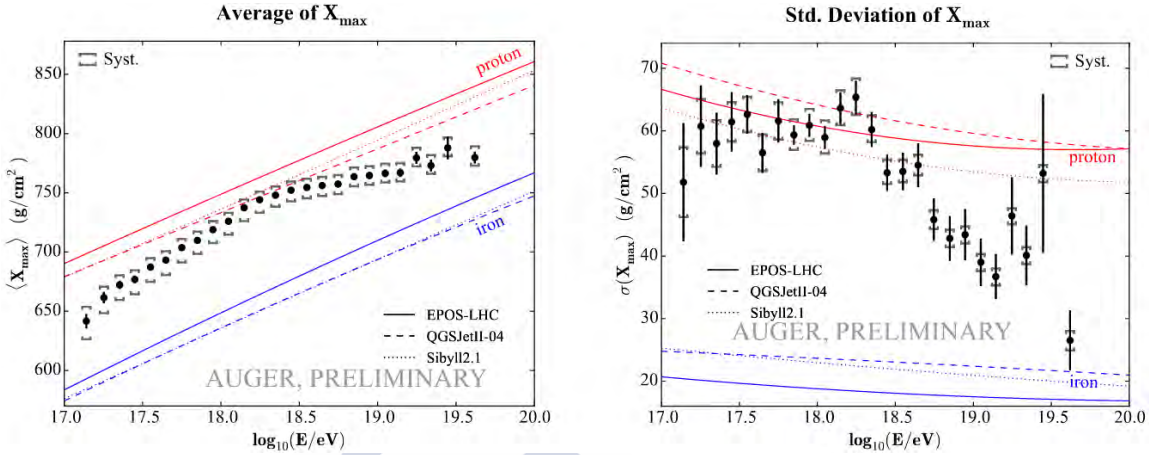
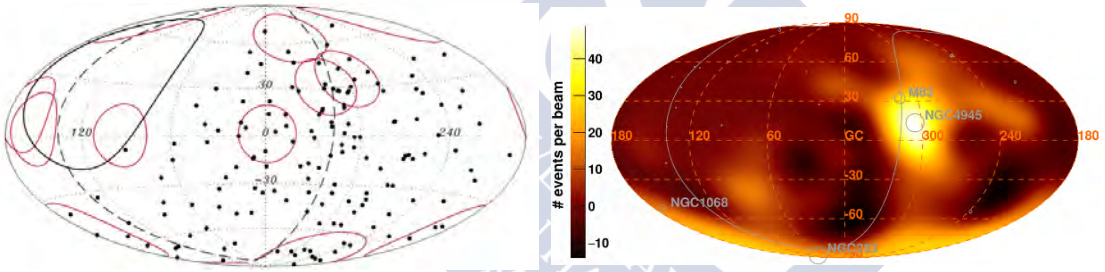


FIGURE 1.9: The mean (left panel) and the standard deviation (right panel) of the measured $X_{\max}$ distributions as a function of the primary energy compared to current EAS simulations for proton and iron primaries (from [43]).

As discussed in SECTION 1.2.1, the most robust EAS observable sensitive to the mass of the primary particle is the depth of maximum of the shower development, $X_{\max}$, directly measured from the longitudinal profile reconstructed with the FD of hybrid events measured simultaneously with both the FD and (at least) one detector of the SD array. The first two central moments of the measured distribution of $X_{\max}$ derived from the combination of the standard FD and HEAT data sets are shown in FIGURE 1.9. The results indicate that the mean primary mass is becoming lighter all the way from 10^{17} to $\approx 10^{18.3}$ eV. Above this energy, the trend inverts and the composition becomes heavier towards the suppression region.

The distribution of arrival directions of UHECRs with energies above $\simeq 4{\times}10^{19}$ eV could reflect the inhomogeneities in the distribution of the nearby extragalactic sources, i.e., those within the local Universe up to few hundred Mpc. The sources should be relatively close due to the energy losses on the CMB and EBL fields (as discussed in SECTION 1.1.2). Despite the low flux of particles in this energy range, the huge collecting area of the Auger Observatory together with its wide field of view from $-90°$ to $+45°$ in declination offer the possibility to search for anisotropy at small and intermediate angular scales at the highest energies with unprecedented statistics what is most valuable to infer the sources of UHECRs. Data have been

subjected to different searches for anisotropies [44], finding the two largest deviations from isotropy for an energy threshold of $58{\times}10^{18}$ eV when looking into the region of the sky within 15° from the location of Centaurus A (the closest radio-loud AGN) and when considering the cross-correlation with the ten most luminous AGNs within 130 Mpc detected in X-ray from the *Swift*-BAT catalogue (see FIGURE 1.10, A). The significance of these findings is not high enough to claim a discovery, yet these excesses are carefully monitored. Recently the Auger collaboration has examined the correlation of the highest energy events with two populations of extragalactic sources of gamma rays, namely Starbursts and AGNs observed by the *Fermi*-LAT satellite above 100 MeV [41]. The result is that for starburst galaxies a 4σ excess is found at an intermediate angular scale of $\sim 13°$ (see FIGURE 1.10, B), while for gamma-ray AGNs the excess is of 2.7σ at an angular scale of $\sim 7°$. If the sources are confirmed the knowledge of the composition will be crucial to understand the propagation of the cosmic rays and could open new areas of study in the Pierre Auger Observatory as the study of the magnetic fields filling the medium between the sources and the Earth.



(A) (B)

FIGURE 1.10: (A) The sky map (in Galactic coordinates) shows the events with $E \geq 58$ EeV together with the *Swift* AGNs brighter than 10^{44}erg s-1 and closer than 130 Mpc, indicated with circles of 18° radius. (B) Observed excess map of events with $E \geq 39$ EeV obtained with starburst. The starburst galaxies with the largest excess weights are indicated.

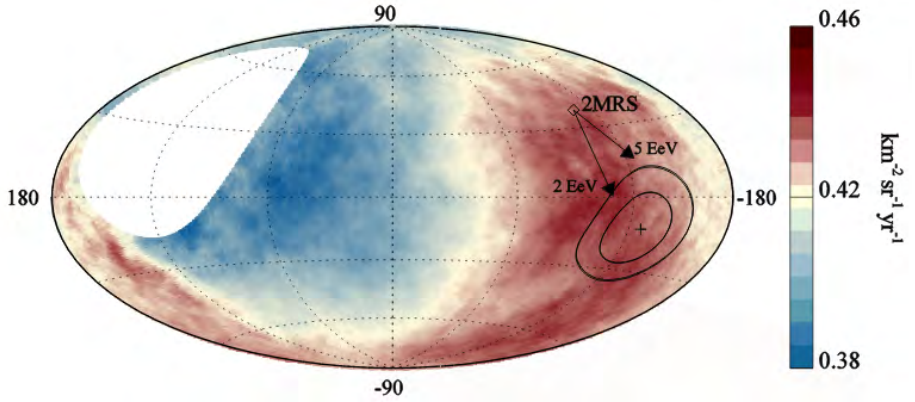


FIGURE 1.11: The sky map (in Galactic coordinates) shows the cosmic-ray flux for $E \geq 8$ EeV smoothed with a 45° top-hat function. The Galactic centre is at the origin. The cross indicates the measured dipole direction. The dipole in the 2MRS galaxy distribution is indicated, while arrows show the deflections expected for a particular model of Galactic magnetic field (see [39] for details).

Large-scale (LS) anisotropies are possible signatures of a collective motion of cosmic rays and/or of the global distribution of their sources at all energies, or of both. The Auger collaboration using 30,000 cosmic rays above $8{\times}10^{18}$ eV has just reported [39] the observation of a LS anisotropy in their arrival directions, detected at more than the 5.2σ level of significance. This can can described by a dipole with an amplitude of $6.5^{+1.3}_{-0.9}\%$ towards right ascension $\alpha_d = (100 \pm 10)^\circ$ and declination $\delta_d = (-24^{+12}_{-13})^\circ$, which is depicted in FIGURE 1.11. By comparing this results with phenomenological predictions, the measured anisotropy supports the hypothesis of an extragalactic origin for the highest-energy cosmic rays.

The Pierre Auger Observatory has also been revealed as a powerful tool to study hadronic interactions at the higher energies, unreachable at current accelerator experiments. The Auger Observatory thus presents an unique window to test the current hadronic interaction models through studies as that based on the measurement and analysis of the muon number in highly inclined showers at ground, R_μ [45], taking advantage of the fact that this type of showers are dominated by muons. The power of such observable as mass tracer was already pointed out in SECTION 1.2.1, but in addition detailed simulations have shown further dependencies of the muon number on hadronic-interaction properties. Comparing the Auger measurements with predictions from hadronic interaction models, a muon deficit in simulations is found to be of order of 30 to $80^{+17}_{-20}\%$ at 10^{19} eV (depending on the model) as shown in FIGURE 1.12. This result allows to claim that the current extrapolations to ultra-high energy interactions must be revisited. Although to fully explore the potential of the muon

number measured at ground to infer the mass composition is needed to resolve the apparent muon deficit in EAS simulations, the logarithmic gain of muons with rising energies gives some insights into composition trend with energy between 4×10^{18} and 5×10^{19} eV. Both $X_{\max}$ and $\langle R_\mu \rangle$ measurements are pointing out a transition from lighter to heavier elements as the energy increases in the considered energy range.

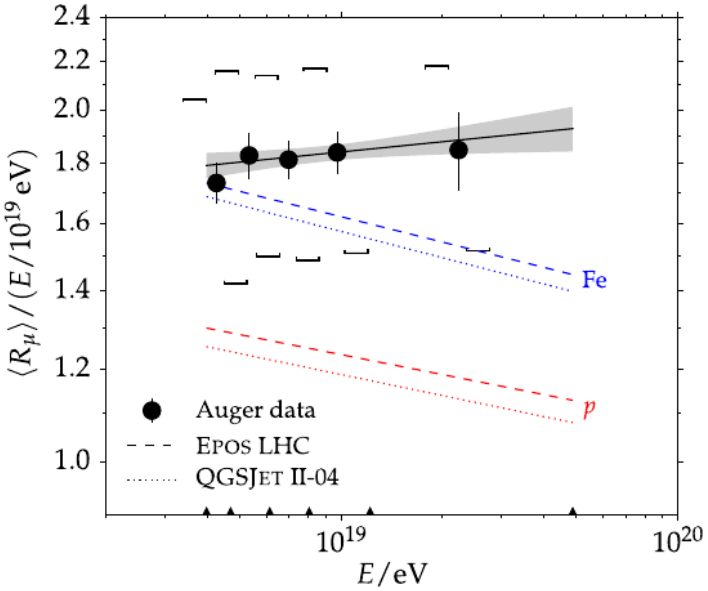


FIGURE 1.12: Average muon content R_μ per shower energy E as a function of the shower energy E. The grey band indicates the statistical uncertainty of the fitted line and square brackets indicate the systematic uncertainty of the measurement. Figure taken from [45]

Although the SD is not separately sensitive to the muonic and EM components of the shower, the measurement of the time structure of the signals from the water-Cherenkov detectors provides observables that can be used to extract information about the development of air showers, and thus to indirectly infer details about the mass composition of the cosmic rays and to prove hadronic interaction models. The Auger collaboration has explored successfully the potential of several parameters based on the signal time-profile observables. One of these parameters related with the production of the muons along the shower is $X^{\mu}_{\max}$ [46]. Others observables directly related with the rise time (the rise time for a single tank is the time for the signal to increase from the 10% to 50% of the total integrated signal) are its azimutal asymmetry [47] and the Δ_s [48]. From comparison of these observables with predictions from hadronic models it is argued that the models are inadequate to describe the various measurements [36, 45–48]. As a consequence the mass composition

inference depends on the level at which the observable relies on the muonic or the electromagnetic component of the shower. These findings suggest that the current hadronic interaction models must be improved to infer in accurate the mass of cosmic rays detected with the Pierre Auger Observatory.

1.5 The place of this thesis

This thesis is aimed to infer the composition of the UHECRs using data recorded at the Pierre Auger Observatory. For this purpose we choose the $X_{\max}$ distributions as mass tracer and we develop new analysis method within the Pierre Auger Collaboration using Bayesian statistics.

We will show the inferences obtained assuming different primary mass scenarios and three different hadronic interaction models: EPOS LHC [49], QGSJETII-04 [50] and SIBYLL 2.1 [51].

As it has been pointed out in SECTION 1.4, the inferences obtained in this work may be subject to criticism because outcomes from different variables lead to different conclusions due to none of the current hadronic models describe all observables of the extensive air showers satisfactorily. The results and conclusions could change in the future as the hadronic models change. Nevertheless, the procedure used along this work will be the same laying the foundations of the Bayesian composition analyses.

Further steps of this work would be to consider other variables, compare the results and also use more than one variable at the same time by performing a multivariate analysis.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 2

Introduction to Bayesian statistical inference

In this chapter some formal statistical background is presented which leads to a brief introduction to Bayesian statistical inference.

2.1 Introduction to probability

The formalisation of probability starts with the definition of the *sample space* $\mathcal{S}$ which is the set containing all possible results of a given experiment. Note that an experiment in this context could be to measure the speed of light, to roll a dice or to predict tomorrow's temperature. When you throw a dice one can ask what is the probability of getting a 2, but also what is the possibility of getting an even number, a prime number or a number greater than 4. Each of these corresponds to an event. The *event space* $\mathcal{A}$ must be defined to include all the possibilities. $\mathcal{A}$ is the set containing all possible events of the experiment, i.e., is the set of all subsets of $\mathcal{S}$, $\mathcal{A} = \mathcal{P}(\mathcal{S})$ (the power set of $\mathcal{S}$). Once the sample space is defined for the experiment (and thus the event space) we need to define a *measure of probability* P which gives a probability $P(A)$ for each event $A \in \mathcal{A}$. The sample space, the event space and the probability $(\mathcal{S}, \mathcal{A}, P)$ are the three basic elements of the probability calculus and constitute the probabilistic space.

2.1.1 The measure of probability P

Kolmogorov gave a formal definition of probability in order to develop a mathematic well-formulated theory of probability. Probability P is an application

$$\begin{array}{rccl} P: & \mathcal{A} & \longrightarrow & \mathbb{R} \\ & A & \longmapsto & P(A) \in [0,1], \end{array} \tag{2.1}$$

i.e, each event of the event space has a probability which is a real number within the interval $[0, 1]$. It must satisfy the following axioms:

- Axiom 1: $\forall A \in \mathcal{A}$ with $\mathcal{A}$ the event space, then:

$$P(A) \geq 0. \tag{2.2}$$

  For each event of the sample space, its probability can not be negative.

- Axiom 2: let $\mathcal{S}$ the sample space, then:

$$P(\mathcal{S}) = 1 \tag{2.3}$$

  This axiom is also called *the assumption of unit measure*: there are no events outside the sample space.

- Axiom 3: $\forall A_i \in \mathcal{A}\ / A_i \cap A_j = \emptyset$ if $i \neq j$ then:

$$P\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} P(A_i) \tag{2.4}$$

  Note that $A_i \cap A_j = \emptyset$ means that the events A_i, A_j are *mutually exclusive* or *disjoint events*. For instance, if the experiment consists on rolling a dice, A_i represents to get a even number greater than 2 and A_j represents to get a prime number, then $A_i \cap A_j$ is the empty set and the events are mutually exclusive.

Note that this definition of probability does not help to establish the probability of the events. Following these axioms the following properties are easy to prove. They are given for completitude:

i) The probability of the empty set is zero: $P(\emptyset) = 0$

ii) If all A_i are mutually exclusive then the $P\left(\bigcup_{i=0}^{n} A_i\right) = \sum_{i=0}^{n} P(A_i)$
$\forall A_i \in \mathcal{A} \; / A_i \cap A_j = \emptyset$ if $i \neq j$ then $P\left(\bigcup_{i=0}^{n} A_i\right) = \sum_{i=0}^{n} P(A_i)$

iii) If A^c and A are complementary events, their union is the sample space and the probability of $P(A^c) = 1 - P(A)$
Let $A^c \in \mathcal{A} \; / A \cap A^c = \emptyset$ and $A \cup A^c = \mathcal{S}$ then $P(A^c) = 1 - P(A)$

iv) If A is a subset of B $(A \subset B)$ then $P(A) \leq P(B)$

v) The probability of an event is always between 0 and 1:
$\forall A \in \mathcal{A} \; 0 \leq P(A) \leq 1$

vi) Let $A, B \in \mathcal{A} / A \cap B \neq 0$, then the probability of the union of two events is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

vii) Let $\{A_i\}_{i=1}^{n} \subset \mathcal{A}$ a set of events not necessarily disjoint, then $P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$

2.1.2 Probability assignment

Using Kolmogorov's axioms and the properties derived from them, it is enough to know the probability of the elements of $\mathcal{S}$ to get the probability of any event of $\mathcal{A}$. All this constitutes a consistent formalism for the probability, however it does not explain how to assign the probability to the elements of the sample space. Any kind of probability law that fulfils Kolmogorov's axioms is valid. Often the Laplace criterion is followed. This criterion could be paraphrased as: *in the absence of any further information (prior information) all possible results should be considered equally probable.* Note that the criterion says "in the absence of further information".

2.1.3 Conditional probability

Let $(\mathcal{S}, \mathcal{A}, P)$ be the probabilistic space and let $A, B \in \mathcal{A}$ be two events with $P(A),\ P(B) \geq 0$. Then the conditional probability of the event A given the event B, i.e, the probability A if B occurs is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.5}$$

where $P(A \cap B)$ is the probability of getting both A and B. It is easy to see that $P(A \cap B) = P(A|B)P(B)$ and this relation is called the multiplication rule. If we have three events A, B and C of the $\mathcal{A}$, the multiplication rule is

$$P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C) \tag{2.6}$$

In general, let $\{A_i\}_{i=1}^n$ a set of events, then the multiplication rule can be generalised to:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{j=1}^{n-1} P\left(A_j | \bigcap_{k=j+1}^{n} A_k\right) P(A_n). \tag{2.7}$$

Sometimes B occurs but this does not provide information on the probability of the occurrence of A, *i.e*, $P(A|B) = P(A)$, then A and B are independent events and the multiplication rule is transformed to the well known result:

$$P(A \cap B) = P(A)P(B). \tag{2.8}$$

2.1.4 Law of total probability

Let $(\mathcal{S}, \mathcal{A}, P)$ the probabilistic space and $\{A_i\}_{i=1}^n$ a partition[1] of $\mathcal{S}$, then $\forall B \in \mathcal{A}$ its probability is

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i), \tag{2.9}$$

This is easy to prove: since $\{A_i\}_{i=1}^n$ is a partition of $\mathcal{S}$, it is possible to write B as:

$$B = \bigcup_{i=1}^{n} (B \cap A_i), \tag{2.10}$$

where $(B \cap A_i) \cap (B \cap A_j) = \emptyset$ if $i \neq j$. Using the second property of the probability derived from Kolmogorov's axioms and the multiplication rule (Equation 2.7):

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{2.11}$$

[1]Given the set C, $\{C_i\}_{i=1}^n$ is a partition of C if $C_i \cap C_j = \emptyset \ \forall i \neq j$ and $\bigcup_{i=1}^n C_i = C$

2.2 Bayes' theorem

The proof of the Bayes' theorem is just an application of the conditional probability and the law of total probability discussed in SECTIONS 2.5-2.1.4. Let $(\mathcal{S}, \mathcal{A}, P)$ be the probabilistic space and $\{A_i\}_{i=1}^n$ a partition of $\mathcal{S}$. Let B an event which is known to occur. The probability of occurrence of the event A_j given the event B is:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)} \tag{2.12}$$

The interpretation of Bayes' formula is very intuitive: if the possible events that can occur are classified in A_i with probabilities $P(A_i)$ (prior probabilities) and the event B is observed, then the Bayes' theorem gives the probabilities modified by the occurrence of B (posterior probabilities).

2.3 Random variables, probability mass functions and probability density functions

A random variable X is a function from the sample space $\mathcal{S}$ to the set of real numbers $\mathbb{R}$.

$$\begin{array}{rccc} X: & \mathcal{S} & \longrightarrow & \mathbb{R} \\ & s & \longmapsto & X(s) \subset \mathbb{R} \end{array} \tag{2.13}$$

Depending on the sample space being discrete or continuous our measure we have to modify our measure of probability. For instance, suppose a cosmic ray with energy 10^{18} eV arriving to the Earth interacts with the atmosphere and produces an extensive air shower.

When we are concerned about the number of muons reaching the ground we our sample space $\mathcal{S}$ is any number of muons that can reach the ground. In this case the most straightforward random variable to choose is the number of muons itself:

$$\begin{array}{rccc} N_\mu: & \mathcal{S} & \longrightarrow & \mathbb{R} \\ & s & \longmapsto & N_\mu(s) = n \in \mathbb{N} \subset \mathbb{R} \end{array} \tag{2.14}$$

Since the number of muons reaching the ground is discrete, the probabilities determining how often a given number of muons, n, reach the ground, are determined by a function $f_{N_\mu}(n)$ called probability mass function (p.m.f). This function gives a

real number for any element of the sample space which describes the probability that exactly n muons reach the ground.

According to the second and third axioms of SECTION 2.1.1 the sum of probabilities of all possible number of muons reaching the ground must be one, which implies:

$$\sum_n f_{N_\mu}(n) = 1. \tag{2.15}$$

Usually, the probability mass function of the number of muons arriving to ground, $f(n)$ is called the distribution of the number of muons reaching the ground.

Using the same cosmic ray example, another random variable could be the atmospheric depth at which the cascade initiated by the nucleus has the maximum number of particles:

$$\begin{aligned} X_{\max} : \ \mathcal{S} &\longrightarrow \mathbb{R} \\ s &\longmapsto X_{\max}(s) = x \subset \mathbb{R} \end{aligned} \tag{2.16}$$

In this case there is a continuous distribution of possible values of depths of maximum and the distribution of probabilities becomes a continuous function of x, which is called probability density function (p.d.f). Its interpretation as a probability requires multiplication by an small interval of depths (see SECTION 2.3.1). The condition that all probabilities must sum one implies:

$$\int_{-\infty}^{\infty} f_{X_{\max}}(x)dx = 1. \tag{2.17}$$

2.3.1 Probability calculus

Let X be a discrete random variable with p.m.f $f_X(x)$. The probability that $X = x_i$ is given by

$$P(X = x_i) = f_X(x_i) \tag{2.18}$$

One can easily get the probability that $X \in [x_i, x_j]$:

$$P(x_i \leq X \leq x_j) = \sum_{k=i}^{j} f_X(x_k) \tag{2.19}$$

If x_0 is the minimum value that X can take, an interesting function is the cumulative distribution function (c.d.f) given by

$$F_X(x) = P(X \leq x) = \sum_{x'=x_0}^{x} f_X(x') \tag{2.20}$$

The c.d.f gives the probability that $X \leq x \ \forall x$ and by definition $F_X(x) \in [0, 1]$. When there is no ambiguity it is customary to denote $F_X(x)$ as $F(x)$ and $f_X(x)$ as $f(x)$. If X is a continuous random variable (rather than a discrete one) with a p.d.f $f_X(x)$, then $P(X = x) = 0$ (this is another way to define a continuous random variable) but we can compute the probability that $a \leq X \leq b$ as

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx \tag{2.21}$$

If $b = c - \epsilon$ and $a = c + \epsilon$ then $P(a \leq X \leq b) = \int_{c-\epsilon}^{c+\epsilon} f_X(x)dx \approx 2\epsilon f(c)$ and the p.d.f evaluated at c can be interpreted as a probability of a random variable in a small interval around c. If x_0 is the minimum value that X can take, the c.d.f of continuous random variables takes the form:

$$F_X(x) = P(X \leq x) = \int_{x'=x_0}^{x} f_X(x')dx' \tag{2.22}$$

Note that knowing the c.d.f implies that the p.d.f (or p.m.f if the random variable is discrete) can be found. Suppose that X is discrete, then

$$f(x) = P(X = x) = F(x) - F(x^-), \tag{2.23}$$

where $F(x^-)$ denotes the cumulative distribution for $X < x$. If X is continuous

$$f(x) = \frac{dF(x)}{dx} \tag{2.24}$$

In the future we going to work with continuous random variables but the results can be easily extrapolated to discrete random variables exchanging integrals by sums.

2.4 Mixture distributions

Let $\{G_i(x)\}_{i=1}^n$ be a set of c.d.f for a random variable and $\{\alpha_i\}_{i=1}^n$ a set of positive real numbers fulfilling $\sum_{i=1}^n \alpha_i = 1$. Then the function

$$F(x) = \sum_{i=1}^{n} \alpha_i G_i(x) \tag{2.25}$$

is also a c.d.f and the p.d.f can be calculated as

$$f(\vec{x}) = \frac{dF(x)}{dx} = \sum_{i=1}^{n} \alpha_i g_i(x) \tag{2.26}$$

This distributions are useful, for example, to describe the effects of a mixture of primaries in cosmic ray data.We can consider the distributions $g_i(x)$ to be the $X_{\max}$ distributions for each primary particle i and then α_i are the relative abundances of the primaries.

2.4.1 Moments of the mixture distributions

Let $h(x)$ some function of the variable X and $f(x) = \sum_{i=1}^n \alpha_i g_i(x)$ an univariate mixture p.d.f. Suppose that X can take values from $-\infty$ to ∞. The expected value of $h(x)$ is given by

$$\begin{aligned} E[h(x)] &= \int_{-\infty}^{\infty} h(x) f(x) dx \\ &= \int_{-\infty}^{\infty} h(x) \left(\sum_{i=1}^n \alpha_i g_i(x)\right) dx \\ &= \sum_{i=1}^n \alpha_i \int_{-\infty}^{\infty} h(x) g_i(x) dx \\ &= \sum_{i=1}^n \alpha_i E_i[h(x)] \end{aligned} \tag{2.27}$$

- m^{th} moment about zero μ'_m: let $h(x) = x^m$, then

$$E[x^m] = \sum_{i=1}^{n} E_i[x^m] = \sum_{i=1}^{n} \alpha_i \mu'_{m,i} \tag{2.28}$$

  where $\mu'_{m,i}$ denotes the m^{th} moment about zero when the values of X are described with the p.d.f $g_i(x)$. In particular the expected value or mean is the first moment about zero:

$$\mu = \mu'_1 = \sum_{i=1}^{n} \alpha_i \mu_i \tag{2.29}$$

- m^{th} central moment μ_m: let $h(x) = (x-\mu)^m$, then

$$\begin{aligned}
\mu_m &= E[(x-\mu)^m] \\
&= \textstyle\sum_{i=1}^n \alpha_i E_i[(x-\mu)^m] \\
&= \textstyle\sum_{i=1}^n \alpha_i E_i[(x-\mu_i+\mu_i-\mu)^m] \\
&= \textstyle\sum_{i=1}^n \alpha_i E_i[\sum_{k=0}^m \binom{m}{k}(x-\mu_i)^k(\mu_i-\mu)^{m-k}] \\
&= \textstyle\sum_{i=1}^n \alpha_i \sum_{k=0}^m \binom{m}{k} E_i[(x-\mu_i)^k(\mu_i-\mu)^{m-k}] \\
&= \textstyle\sum_{i=1}^n \alpha_i \sum_{k=0}^m (\mu_i-\mu)^{m-k}\mu_{m,i}
\end{aligned} \tag{2.30}$$

The variance is given by the second central moment and for a mixture distribution has the following expression in terms of the mean (μ_i) and variance (σ_i^2) of each component:

$$\sigma^2 = \mu_2 = \sum_{i=1}^n \alpha_i[(\mu_i-\mu)^2 + \sigma_i^2] \tag{2.31}$$

2.5 Joint density functions

We can easily extend the definitions given before to two or more random variables. Let X and Y two random variables. The probability of $a \leq X \leq b$ and $c \leq Y \leq d$ is given by

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)dydx \tag{2.32}$$

where $f_{X,Y}(x,y)$ is called the joint p.d.f. The probability of $a \leq X \leq b$ for all possible values of Y is

$$P(a \leq X \leq b) = \int_a^b \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx = \int_a^b f_X(x)dx \tag{2.33}$$

where $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ is called the *marginal p.d.f of* X. In the same way the marginal p.d.f of Y is defined as $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$. The conditional probability of X given a value of Y can be derived from (2.5) as

$$\begin{aligned}
P(a \leq X \leq b | y-\epsilon \leq Y \leq y+\epsilon) &= \tfrac{P(a \leq X \leq b, y-\epsilon \leq Y \leq y+\epsilon)}{P(y-\epsilon \leq Y \leq y+\epsilon)} \\
&= \tfrac{\int_a^b \int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(x,y')dy'dx}{\int_{-\infty}^{\infty} \int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(x,y')dy'dx} \\
&\approx \tfrac{\int_a^b f_{X,Y}(x,y)dx}{f_Y(y)}
\end{aligned} \tag{2.34}$$

We arrive to the definition of conditional p.d.f as:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{2.35}$$

Note that a Bayes' theorem in terms of probability density functions can be obtained in the same way as in SECTION 2.2:

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x) \Leftrightarrow f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}. \tag{2.36}$$

Here, $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$ denote the conditional probability density functions. Once more, one usually writes $f(y|x)$ instead of $f_{Y|X}(y|x)$. This concept is easily generalised to more than two variables (as it has been shown in SECTION 2.1.3). For instance, if we have n variables we can write:

$$f(x_1, ..., x_n) = f(x_1|x_2, ..., x_n) \cdots f(x_n), \tag{2.37}$$

where the calculation of $f(x_n)$ requires the integration over the other $n-1$ variables.

Now all statistical concepts necessary to understand the Bayesian inference have been exposed.

2.6 Bayesian inference

Let $D = \{x_i\}_{i=1}^n$ be n realisations of a random variable X, *i.e,* n results of experiments consisting in measuring the variable X. Let θ be a parameter of interest. The Bayesian inference consists of allocating probabilities to the possible values of θ accordingly to the observed data set D by solving the equation

$$\pi(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{f(D)} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} \tag{2.38}$$

In the following the meaning of the individual terms of EQUATION 2.38 are explained.

2.6.1 The likelihood function

The likelihood function $f(D|\theta)$ is the conditional probability distribution of D given the unknown parameter θ and it is usually denoted as $\mathcal{L}(\theta|D)$. This function describes

how the data set D is distributed assuming a given value of θ and is called *likelihood function*. Given D the value of θ is more likely to be the true parameter than θ' if $\mathcal{L}(\theta|D) > \mathcal{L}(\theta'|D)$. Assuming that the different measurements of the data set are independent, the likelihood function can be written as

$$\mathcal{L}(\theta|D) = \prod_{i=1}^{n} f(x_i|\theta) \tag{2.39}$$

The maximum likelihood estimation consists on the estimation of the true value of θ maximising the likelihood function. This satisfies the likelihood principle: *All the information about θ obtainable from an experiment is contained in the likelihood function for θ given X. Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to one another*, see [52] and [53].

The likelihood principle is not accepted by all the scientists, in fact, it is not accepted by the "Frequentist" because different inferences about θ can be derived with two proportional likelihoods using the Frequentist approach. In [53] a discussion is presented this topic and it is also shown that the likelihood principle is derived by the assumption of two principles: the principle of sufficiency and the principle of conditionality, which are paradoxically accepted by most of the scientific community. These principles can be described informally as asserting the "irrelevance of observations independent of a sufficient statistic" (sufficiency) and the "irrelevance of experiments not actually performed" (conditionality).

2.6.2 The prior

The other term in the numerator of the right side of Equation 2.38 is the prior probability density function of θ, $\pi(\theta)$, often called prior. This function describes all the information that we have about the parameter of interest before performing the experiment. A prior distribution can be created using information about past experiments, using theoretical knowledge or expressing our total ignorance about the problem. For example, in the problem of inferring the primary composition of cosmic rays reaching the Earth with energies above 1 EeV, a possible prior choice could be describing a proton dominance because protons are the most abundant nuclei at low energies. Another prior choice could be describing heavier elements than proton because a given the theoretical model says that the nuclei are not produced

with the observed energies but are accelerated in magnetic fields, and then, they are accelerated accordingly with their electric charge. In fact, the two mentioned priors are equally acceptable and configure two different scenarios or models. In SECTION 2.6.4 the model selection is explained. Often there is no clear a priori knowledge about θ. In this case a uniform prior is assumed meaning that all values of θ are completely equivalent (Laplace criterion rule, SECTION 2.1.2).

2.6.3 The posterior

There are two remaining terms in EQUATION 2.38 to be explained. One is $f(D)$ and the other one is $\pi(\theta|D)$. For now $f(D) \in \mathbb{R}$ is just a normalisation constant ensuring that the posterior probability density function (or simply posterior) has a unit integral: $\int \pi(\theta|D)d\theta = 1$. The posterior $\pi(\theta|D)$ describes our knowledge about the θ parameter after the data analysis of the experimental results. Then one can read EQUATION 2.38 as an update of the prior knowledge of θ, described by the prior, through the experiment described by the likelihood. For each event $x_i \in D$ of the data set, our knowledge about θ changes. Once the posterior distribution is known there are two standard estimators for the true value of θ: the mean of the posterior and the mode (the so called **M**aximum of **A** **P**osteriori distribution, MAP). There is not a simple rule to use one or another and sometimes the choice depends on the posterior distribution itself. For instance in FIGURE 2.1 an example of a bimodal posterior distribution is shown. It is clear that θ can be around 2 or 8 but not around 5 which is the mean value of the posterior distribution. In the next chapter the best estimator for the composition analysis is studied in detail.

It is usually not enough only to estimate the parameter, but it is also necessary to give a range in which the parameter of interest can take values with a certain probability q. This range is called probability interval, confidence interval or credible set. In the Bayesian approach these sets are very easy to calculate. Suppose that the posterior distribution $\pi(\theta|D)$ is known and one want to find between which values $[\theta_1, \theta_2]$ the actual value of the parameter has been estimated. Usually this question is answered with an associated probability q which is typically 0.68, 0.9 and 0.95. The limits of the range are given by solving the equation

$$q = P(\theta_{low} \leq \theta \leq \theta_{up}) = \int_{\theta_1}^{\theta_2} \pi(\theta|D)d\theta \tag{2.40}$$

When the maximum of the posterior distribution is equal or near to one of the limits of the Θ space ($\theta \in \Theta$), one can solve the EQUATION 2.40 as follows:

$$q = \int_{\theta_{min}}^{\theta_1} \pi(\theta|D)d\theta = \Pi(\theta_1|D) - \Pi(\theta_{min}|D) \tag{2.41}$$

$$q = \int_{\theta_2}^{\theta_{max}} \pi(\theta|D)d\theta = \Pi(\theta_{max}|D) - \Pi(\theta_2|D) \tag{2.42}$$

where $\Pi(\theta|D)$ is the posterior cumulative distribution. If the mode of the posterior distribution is nearest to θ_{min} (θ_{max}) one can use EQUATION 2.42 (EQUATION 2.41) to calculate an upper (lower) limit at q of probability, *i.e*, the inferred value of θ is smaller (higher) than the upper (lower) limit with a probability of q. In the next chapter this procedure is compared with the Frequentist approach in the cosmic ray composition inference.

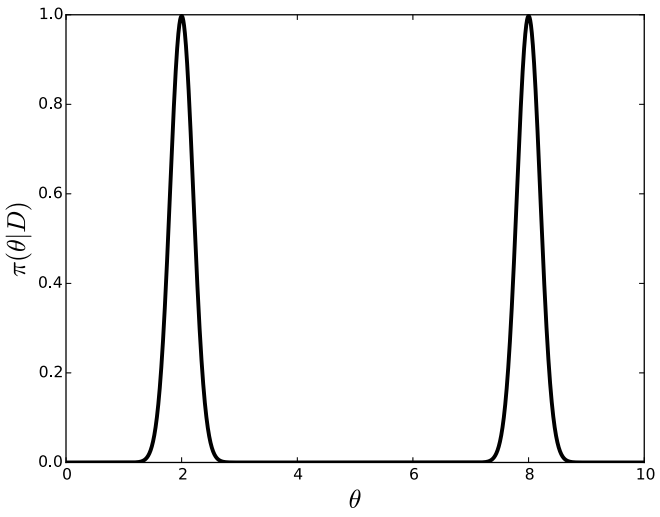


FIGURE 2.1: Example of a bimodal posterior distribution.

2.6.4 The evidence: Bayes' factors and model selection

The denominator in EQUATION 2.38, $f(D)$ is called *evidence* and usually denoted by Z. Up to now the evidence has been considered as a normalisation constant but it takes an important role when comparing different scenarios, models or hypotheses. Consider now two hypotheses H_1 and H_2 that we want to contrast and we perform

an experiment which gives us the data set $D = \{x_i\}_{i=1}^n$. We are going to consider that the likelihood functions are different for the different hypotheses, for H_1 we have $\mathcal{L}_1(\theta|D) = f_1(D|\theta)$ and for H_2 we have $\mathcal{L}_2(\omega|D) = f_2(D|\omega)$ where θ and ω could in principle have different dimensions (θ could be for instance a shape of an exponential distribution and ω could be the mean and the variance of a normal distribution). The posterior distributions are given by

$$\pi(\theta|D, H_1) = \frac{\mathcal{L}_1(\theta|D)\pi(\theta|H_1)}{Z_1} \tag{2.43}$$

for the first hypothesis and

$$\pi(\omega|D, H_2) = \frac{\mathcal{L}_2(\omega|D)\pi(\omega|H_2)}{Z_2} \tag{2.44}$$

for the second hypothesis. Z_1 and Z_2 are the normalisation factors for their respective equations:

$$Z_1 = \int f_1(D|\theta)\pi(\theta|H_1)d\theta = P(D|H_1), \tag{2.45}$$

which gives the probability of the data set D given the hypothesis H_1 (once $P(D|H_k)$ has been normalised to all the hypotheses). In the same way, Z_2 is the probability of D given the hypothesis H_2. The evidences have statistical meaning. Since we can calculate $P(D|H_1)$ and $P(D|H_2)$ we can also calculate $P(H_1|D)$ and $P(H_2|D)$ using the Bayes' theorem obtaining the probability of a given hypothesis given the data set D and independently of the parameters θ and ω:

$$P(H_m|D) = \frac{P(D|H_m)P(H_m)}{P(D)} = \frac{Z_m P(H_m)}{\sum_{l=1}^{M} Z_l P(H_l)}, \tag{2.46}$$

where here $M = 2$ and $m = 1, 2$. The expression shown in Equation 2.46 is the generalisation for M possible hypotheses.

Once more the prior probabilities $P(H_1)$ and $P(H_2)$ must be chosen before the analysis. In this way, we obtain a probability mass function in which the variables are the different hypotheses. To compare which of the hypotheses is preferred by data, the ratio between the posterior probabilities is performed:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{Z_1}{Z_2}\frac{P(H_1)}{P(H_2)}. \tag{2.47}$$

This ratio is called "posterior odds" and the ratio $\frac{P(H_1)}{P(H_2)}$ is called "prior odds". The ratio of the evidences $\frac{Z_1}{Z_2}$ is called the *Bayes' factor* of the hypothesis H_1 over H_2

($B_{1,2}$) and represents the gain of probability of H_1 over the hypothesis H_2 after the data analysis:

$$\text{posterior odds } (H_1, H_2) = B_{1,2} \times \text{prior odds } (H_1, H_2). \tag{2.48}$$

When the prior probabilities are equal ($P(H_1) = P(H_2) = 1/2$) the posterior odds are just the Bayes' factor, *i.e*, the ratio between the evidences. Note that the likelihood ratio test is a special case of the Bayesian posterior odds: the posterior odds becomes the likelihood ratio test[2] when: the likelihoods have the same parametric form, $P(H_1) = P(H_2)$ and $\pi(\theta|H_i) = \delta(\theta - \theta_i)$, *i.e*, when we identify the hypotheses with given values of θ.

The Bayesian model comparison takes into account our prior knowledge on the hypotheses, the number of parameters of these hypothesis and the analysis of data. As it is shown in Appendix D an important attribute of the Bayesian model selection is that it tries to favour the simplest model.

For a generalised number of models $H_1, H_2, ..., H_M$ the probabilistic interpretation given by Equation 2.46 is still valid and allows us to treat the models as if they were random variables. Let Δ a parameter in which we are interested. For one of the models, the posterior distribution of the parameter given the data D is obtained using the Bayes' theorem

$$\pi(\Delta|D, H_m) = \frac{\mathcal{L}(\Delta|D, H_m)\pi(\Delta|H_m)}{Z_m}. \tag{2.49}$$

For each model we are going to obtain different estimations of the parameter Δ because Equation 2.49 gives us the probability density function of the parameter Δ given the data set D and assuming the model H_m. Nevertheless, since we can calculate the probability of the model given the data (Equation 2.46) we can apply the law of total probability to the $\pi(\Delta|D, H_m)$ distribution:

$$\pi(\Delta|D) = \sum_{m=1}^{M} \pi(\Delta|D, H_m)Z_m. \tag{2.50}$$

Here, $\pi(\Delta|D)$ is the probability density function of the parameter Δ given the data taking into account the uncertainty on the model which represents the data distribution. This method is called *Bayesian Model Averaging* (BMA). See [54] for a review

[2]Usually in the likelihood ratio test one of the hypothesis is considered the *null hypothesis* changing their subscript by zero.

with applications and examples. Even when EQUATION 2.50 has been deduced using probabilistic reasoning it is just a mixture distribution composed by M distributions $\{\pi(\Delta|D, H_m)\}_{m=1}^{M}$ which weights or fractions are $\{Z_m\}_{m=1}^{M}$, then we can obtain easily the expected value ($E[\Delta|D]$) and the variance ($V[\Delta|D]$) of Δ (see SECTION 2.4):

$$E[\Delta|D] = \sum_{m=1}^{M} E[\Delta|D, H_m]Z_m, \tag{2.51}$$

$$V[\Delta|D] = \sum_{m=1} \left\{ \left(E[\Delta|D, H_m] - E[\Delta|D]\right)^2 + V[\Delta|D, H_m] \right\} Z_m, \tag{2.52}$$

where $E[\Delta|D, H_m]$ and $V[\Delta|D, H_m]$ are the expected value and the variance of $\pi(\Delta|D, H_m)$ respectively. As the number of events in the data sample D increases, the evidence in favour of the true model (if it is accounted for in the analysis) also increases so that if H_t is the true model and N is the number of events in the data sample:

$$\lim_{N\to\infty} \pi(\Delta|D) = \pi(\Delta|D, H_t). \tag{2.53}$$

The BMA is a natural way to deal with uncertainties in the model selection.

2.6.5 Predictive distributions

Finally we are going to present *predictive distributions*. Suppose that an observer wants to prepare an experiment to infer certain parameter θ which can take values in the Θ space with prior probabilities $\pi(\theta)$. The distribution of the random variable X is given by the likelihood function $f(x|\theta)$. The data distribution before the experiment is

$$f(\tilde{x}) = \int_{\Theta} f(\tilde{x}|\theta)\pi(\theta)d\theta \tag{2.54}$$

where $\tilde{x}$ denotes unobserved data. $f(\tilde{x})$ is called the prior predictive distribution. After the experiment has been built and the data D analysed, the knowledge about θ has changed: $\pi(\theta) \to \pi(\theta|D)$. Now the expected data distribution has also changed:

$$f(\tilde{x}) \to f(\tilde{x}|D) = \int_{\Theta} f(\tilde{x}|\theta)\pi(\theta|D)d\theta \tag{2.55}$$

where $f(\tilde{x}|D)$ is called the posterior predictive distribution. This distribution can be used for instance if the observer is thinking about an experiment update or needs to compare with the observed data distribution to get a feeling of how well the estimation of θ fits the measured data. In the next chapters the posterior predictive

distribution is used in the second way. Of course, the posterior predictive distribution of one experiment can be used as a prior predictive distribution of another future experiment.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 3

Methods for composition analysis

In this we intend to compare using the most widely used methods the composition analysis fractions in cosmic ray data. These methods are namely the maximum likelihood, the χ^2 minimisation, the mean value of the distributions (method of moments) and the mean value of the posterior probability density function. The discrimination power of the four methods is discussed in three different physical scenarios of increasing complexity: signal to noise discrimination, the inference of the proportions of a sample of a mixture of two distributions and finally the case of interest for this thesis, the $X_{\max}$ distributions for mixed primary mass composition. The determination of the confidence levels in composition analysis is studied in both the Bayesian and Frequentist approaches and the estimation of the composition when the data is biased by different detector effects is addressed in a progressive way to understand the power of Bayes' methods and to compare with alternative approaches such as the "anti-bias" method which is currently used for the analysis of the $X_{\max}$ data of the Pierre Auger Observatory . This chapter does not deal with real data. We have studied the power of these methods using simulations and implementing detector effects in a realistic way. When we refer to data in different sections we will be referring to mock data from simulations. Several of these mock "data sets" will be considered along the chapter.

3.1 Methods

Consider the following problem. The value of a single variable x (e.g. the depth of shower maximum, $X_{\max}$) is extracted from two different probability distributions,

$g_1(x)$ and $g_2(x)$ with a "composition" fraction α ($0 \leq \alpha \leq 1$), so that the joint probability distribution is given by the so called mixture distribution (see SECTION 2.4)

$$f(x|\alpha) = \alpha g_1(x) + (1-\alpha) g_2(x). \tag{3.1}$$

In this example the probability distributions $g_{1,2}(x)$ are known and the problem consists in determining the composition fraction α from the measurement of n data points x_i, $i = 1, \cdots, n$. If α was known, the probability of getting the data set $D = \{x_i\}$ would be given by

$$P(D|\alpha) = \mathcal{L}(\alpha|D) = \prod_{i=1}^{n} f(x_i|\alpha), \tag{3.2}$$

This is called the likelihood of α given the data set D. Here we are implicitly assuming that the different data points are independent. Using Bayes' theorem [55], we can obtain the posterior probability density function for α given the data as

$$P(\alpha|D, I) = \frac{P(D|\alpha, I) P(\alpha|I)}{P(D|I)}. \tag{3.3}$$

Here I is any prior information we have about the problem, including the prescription of the probabilities g_i, for this reason, we have changed $P(D|\alpha)$ by $P(D|\alpha, I)$. $P(D|I)$ is the probability of obtaining the given data independently of any value of α and here acts as a normalisation constant. If $P(\alpha|I)$ is described by the probability density function $\pi(\alpha|I)$ and $P(\alpha|D, I)$ is described by $\pi(\alpha|D, I)$ we can write EQUATION 3.3 in terms of the probability density functions[1] as

$$\pi(\alpha|D, I) = \frac{\mathcal{L}(\alpha|D, I)\pi(\alpha|I)}{\int_0^1 \mathcal{L}(\alpha|I)\pi(\alpha|I)d\alpha}, \tag{3.4}$$

In our problem, information on the cosmic ray composition could arise from astrophysical reasoning and give preference for, say, proton dominance. In the absence of any information a flat distribution on the space of the parameter of interest is reasonable and follows the Laplace criterion. In the following we will use $\pi(\alpha) = Uniform(0, 1)$, but all our results will be equally valid for any other assumptions or choices of prior probabilities.

[1] Here we are approximating the probability around α, $P(\alpha|D)$ as the value of the probability density function in α, $\pi(\alpha|D)$, as was explained in the previous chapter but in fact, $\pi(\alpha|D)$ is a probability density and not a probability.

Therefore, we can write EQUATION 3.4 as

$$\pi(\alpha|D) = \frac{1}{\mathcal{N}} \prod_{i=1}^{N} \{\alpha g_1(x_i) + (1-\alpha) g_2(x_i)\}\,, \tag{3.5}$$

where $\mathcal{N}$ is a normalisation constant and the prior I is omitted for clarity [2]. In some practical situations instead of EQUATION 3.5, where all the data points are given, one can have data binned in the variable x. In that case the equation reads

$$\pi(\alpha|D_k) = \frac{1}{\mathcal{N}} \prod_{j=1}^{k} \{\alpha G_1(x_j) + (1-\alpha) G_2(x_j)\}^{n_k}\,, \tag{3.6}$$

where the data now is $D_k = \{n_1, \cdots, n_k\}$, the number of events in the bins $1, \cdots, k$ with center values $x_1, \cdots, x_k$, and $G_i(x_j) = \int_{x_j} g_i(x)dx$ is the integral over the bin integral of the probability density. Although binning the data makes the problem somehow easier, it wastes information especially in regions where the derivative of the function has a large value.

EQUATION 3.5 (or alternatively EQUATION 3.6) contains all the information we have about the problem. Estimation of the composition fraction simply reduces to the choice of the "best estimator" of α. To answer this question different composition-fraction estimators can be currently found in the literature. Several of these options are discussed in the following.

i) One can calculate the mean of the data points and choose α such that it coincides with the mean value of the distributions, *i.e.* if $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, then

$$\bar{x} = \alpha \langle x \rangle_{g_1} + (1-\alpha) \langle x \rangle_{g_2} \tag{3.7}$$

where $\langle x \rangle_{g_i} = \int x g_i(x) dx$. It is straightforward to get the estimate which will be denoted as $\alpha_{<>}$:

$$\alpha_{<>} = \frac{\bar{x} - \langle x \rangle_{g_2}}{\langle x \rangle_{g_1} - \langle x \rangle_{g_2}}. \tag{3.8}$$

Eq. (3.8) has the advantage that has a simple analytic form and is easy to evaluate for any distribution (provided it has first moments). However this method is not a good option since it can give unphysical results (for instance, $\alpha > 1$ or $\alpha < 0$). Moreover it gives the largest deviation with respect to the true

[2] Sometimes once the prior has been chosen it will be omitted in order to lighten the equations.

value for all studied estimators as will be shown in the next section. In addition, it is useless if the mean values of the two distributions coincide ($\bar{x}_1 = \bar{x}_2$).

ii) Alternatively, one can choose as an estimator the value that maximises the posterior probability density function.

$$\frac{\partial \pi(\alpha|D)}{\partial \alpha} = 0. \tag{3.9}$$

In this example the posterior probability is given by

$$\pi(\alpha|D) \propto \prod_{i=1}^{N} \{\alpha g_1(x_i) + (1-\alpha) g_2(x_i)\}\, \pi(\alpha). \tag{3.10}$$

Since $\pi(\alpha|D)$ is positive an equivalent alternative is to maximise the logarithm of $\pi(\alpha|D)$.

$$\frac{\partial \log \pi(\alpha|D)}{\partial \alpha} = \sum_{i=1}^{n} \frac{g_1(x_i) - g_2(x_i)}{\alpha g_1(x_i) + (1-\alpha) g_2(x_i)} = 0. \tag{3.11}$$

Note that this estimation becomes the maximum likelihood estimation when the prior $\pi(\alpha)$ is flat in the space of α and the search space of α in the maximum likelihood is bounded between 0 and 1. It is known to give a very good estimation of α in almost all cases, even for small number of events. It has the disadvantage that an analytic solution is possible only for very small number of events or bins. This method is used, for instance, in the standard package TFractionFitter [56] (of the data analysis framework ROOT [57]) to fit the fraction of given data histograms. Usually, the solution is found by numerically searching for the solution of Equation 3.9 or Equation 3.11. The inference of α through this method will be denoted α_{max}.

iii) If the number of events is large, one expects a well defined peak distribution in α. Near the maximum of the distribution one can approximate this distribution by a Gaussian. When the data are binned, we can easily construct a χ^2 variable for the problem

$$\chi^2(\alpha) = \sum_{j=1}^{k} \frac{(n_j/n - F(x_j, \alpha))^2}{n_j}, \tag{3.12}$$

where n_j is the number of data events in bin j and $F(x_j, \alpha) = \alpha G_1(x_j) + (1-\alpha) G_2(x_j)$ is the probability of having an event in bin j for a given α and $G_i(x_j) = \int_{x_{min,j}}^{x_{max,j}} g_i(x) dx$ the integral on the bin j of the probability density of each pure composition.

The optimal value of α can be found minimising the χ^2

$$\frac{\partial \chi^2(\alpha)}{\partial \alpha} = 0. \tag{3.13}$$

The solution of EQUATION 3.13 has the advantage of being analytic and having relatively simple expression regardless the number of bins:

$$\alpha_\chi = \frac{\sum_{j=1}^{k}(n_j/n - G_2(x_j))(G_1(x_j) - G_2(x_j))}{\sum_{j=1}^{k}(G_1(x_j) - G_2(x_j))^2}. \tag{3.14}$$

It is an asymptotic limit of the maximum likelihood method (for n and n_j large) and, as could be expected, it gives very good results in this limit.

iv) Once we know the posterior p.d.f $\pi(\alpha|D)$, we can obtain the mean value of the distribution as

$$\langle \alpha \rangle = \int_0^1 d\alpha \; \alpha \; \pi(\alpha|D). \tag{3.15}$$

Although this estimator is not used as often as the other options described above, indeed it will shown below that it gives the best performance in most cases. It has the disadvantage of being difficult to evaluate analytically but for the simplest cases.

v) There are course, other estimators that can provide sensible results. As an example, the median of the posterior probability, defined by

$$\int_0^{\alpha_M} d\alpha \pi(\alpha|D) = \int_{\alpha_M}^1 d\alpha \pi(\alpha|D) = 1/2, \tag{3.16}$$

which is well known to be a robust estimator, being invariant against a large set of transformations of the probability distributions. However, it is also difficult to evaluate both analytically and numerically. Therefore we will not consider it any further.

3.2 A toy analytic case: mixture of two non-overlapping components

One of the simplest problems of discrimination is when the two distributions $g_1(x)$ and $g_2(x)$ are totally separated (*i.e.* they do not overlap). This problem is equivalent

to determining the probability of having heads or tails in a (possibly) loaded coin. The actual shape of g_1 and g_2 is irrelevant and one can bin the data in only two bins $x = a, b$ such that all the probability is concentrated in either bin a or b. So, let the two probability functions be

$$G_1(x) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{if } x = b \end{cases} \tag{3.17}$$

$$G_2(x) = \begin{cases} 0 & \text{if } x = a \\ 1 & \text{if } x = b \end{cases} \tag{3.18}$$

For simplicity the posterior p.d.f will be labelled as $\phi(\alpha) = \pi(\alpha|D)$ in the following (unless otherwise indicated). We performed N trials in which n is the number of events with $x = a$ and $N - n$ is the number of events with $x = b$. The posterior p.d.f is given by

$$\phi(\alpha) = \frac{1}{\mathcal{N}} \left[\alpha^n (1-\alpha)^{N-n}\right]. \tag{3.19}$$

As could be expected, this distribution corresponds to a $Binomial(\alpha)$ in the space of n but it corresponds to a Beta distribution $Beta(n+1, N-n+1)$ in the composition space [3].

By direct calculation, one can obtain the composition estimators described in the previous section. The inferences using the χ^2 minimisation, the maximum likelihood and the averages of the theoretical distributions are exactly identical

$$\alpha_\chi = \alpha_{\max} = \alpha_{<>} = \frac{n}{N}. \tag{3.20}$$

leading to the result of the estimation of the probability of "head" events to be fraction of head events observed. On the other hand the mean of the posterior p.d.f of α gives a different outcome:

$$\langle\alpha\rangle = \frac{n+1}{N+2}. \tag{3.21}$$

Although this finding may be surprising at first sight it is a well known result in the literature. It is known as Laplace's succession rule. One may notice that in the limit

[3] The random variable Y is Beta distributed if its p.d.f is

$$f(y|\gamma_1, \gamma_2) = \frac{1}{B(\gamma_1, \gamma_2)} y^{\gamma_1 - 1}(1-y)^{\gamma_2 - 1},$$

where $B(\gamma_1, \gamma_2)$ is the Euler beta function. In that case its mode is given by $\frac{\gamma_1 - 1}{\gamma_1 + \gamma_2 - 2}$ and its mean value is given by $\frac{\gamma_1}{\gamma_1 + \gamma_2}$

$N, n \to \infty$ with n/N fixed one recovers EQUATION 3.20. Note that if $N = 0$, then $n = 0$ and all the methods are indefinite except $\langle\alpha\rangle$ which gives $\frac{1}{2}$, just the mean value of the prior probability. If $N = 1$, then either $n = 0$ or $n = 1$, which would give either $\alpha_{\max} = 0$ or 1. EQUATION 3.21 gives $\langle\alpha\rangle = 1/3$ or $2/3$. In SECTION 3.3 this phenomenon is shown numerically for a more realistic case.

3.2.1 Mixture of two distributions with contamination

For a more interesting case, consider now the previous example but with a (possibly small) contamination between both distributions described by ϵ and δ:

$$G_1(x) = \begin{cases} 1-\epsilon & \text{if } x = a \\ \epsilon & \text{if } x = b \end{cases} \tag{3.22}$$

$$G_2(x) = \begin{cases} \delta & \text{if } x = a \\ 1-\delta & \text{if } x = b \end{cases} \tag{3.23}$$

So that there is a (small) probability of a event of type 1 ("heads") to be identified in the bin 2 ("tails") and vice-versa. The posterior probability of α after measuring $N = n_1 + n_2$ total events with n_1 of type 1 and n_2 of type 2 is thus

$$\phi(\alpha) = \frac{1}{\mathcal{N}} \left[\alpha(1-\epsilon) + (1-\alpha)\delta\right]^{n_1} \left[\alpha\epsilon + (1-\alpha)(1-\delta)\right]^{n_2}. \tag{3.24}$$

After some algebra one can obtain again identical results for the first three estimators:

$$\alpha_\chi = \alpha_{\max} = \alpha_{<>} = \frac{1}{1-\delta-\epsilon}\left[\frac{n_1}{N} - \delta\right]. \tag{3.25}$$

In this case the mean value of the posterior distribution, α, does not have a simple analytic expression, turning out to be

$$\langle\alpha\rangle = \frac{1}{1-\delta-\epsilon}\left(\frac{B(1-\epsilon, n_1+2, n_2+1) - B(\delta, n_1+2, n_2+1)}{B(1-\epsilon, n_1+1, n_2+1) - B(\delta, n_1+1, n_2+1)}\right) - \frac{\delta}{1-\delta-\epsilon}, \tag{3.26}$$

where $B(x, n_1, n_2)$ is the incomplete Beta function [58]

$$B(x, n_1, n_2) = \int_0^x dy y^{n_1-1}(1-y)^{n_2-1}. \tag{3.27}$$

For n_1 and n_2 integers, B is a polynomial in x. One can show that the above equation gives always physical values $0 \leq \alpha \leq 1$.

Although this model is rather simplistic, it has all the ingredients found in the relevant cases we will address. One can interpret EQUATION 3.25 rather easily, the term $-\delta$ subtracts the expected fraction of events of type 2 which fall into bin 1. On the other hand the factor $1-\delta-\epsilon$ is a measure of the fraction of well identified events. The overlapping of the two distributions is given by $\delta+\epsilon$. As we will see below, the overlapping area of the distributions is a general characteristic of the problem.

Another interesting point of EQUATION 3.25 is the fact that it can produce unphysical results. If $n/N < \delta$, the expected fraction is negative. This is so because even for $\alpha = 0$, we expect a number of events in the first bin of $\delta \times N$. Finally, one can see that the case $\epsilon + \delta = 1$ is ill defined. But in this case both distributions are equal: hence no discrimination can be made between the two distributions. The calculation of the mean value of the posterior p.d.f ,$\langle\alpha\rangle$, does not suffer from this behaviour, always giving physically admissible results. In the last case, when the two distributions are equal, we would obtain $\langle\alpha\rangle = 1/2$, which is easily interpreted. If the data can not differentiate between the two cases we do not gain any information from the data and the estimation given by our prior is kept.

3.3 Application of the methods

The methods discussed previously are now applied to several different scenarios. In SECTION 3.3.1 a typical problem of signal/noise identification is studied. In SECTION 3.3.2, we concentrate on the separation of two signals and the dependence of the resolution with respect to the "distance" between the two signals, which is a measure that reflects how well we can distinguish between the two distributions.

3.3.1 Signal/Noise discrimination

Consider the case of extracting a signal with a well defined peak from events coming from the signal plus a flat noise. To be concrete, we will choose the following probability density functions

$$g_1(x) = \frac{1}{\mathcal{N}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}; x \in [a,b], \tag{3.28}$$

$$g_2(x) = \frac{1}{b-a}; x \in [a,b]. \tag{3.29}$$

Here $[a, b]$ is the range of the variable ,μ and σ are the mean and the standard deviation of the Gaussian and $\mathcal{N}$ is the appropriate normalisation constant. In the numerical calculations we will choose $a = 0$, $b = 7$, $\Delta = 1$, $\mu = 2$, and $\sigma = 0.2$.

In FIGURE 3.1 we show both probability density functions.

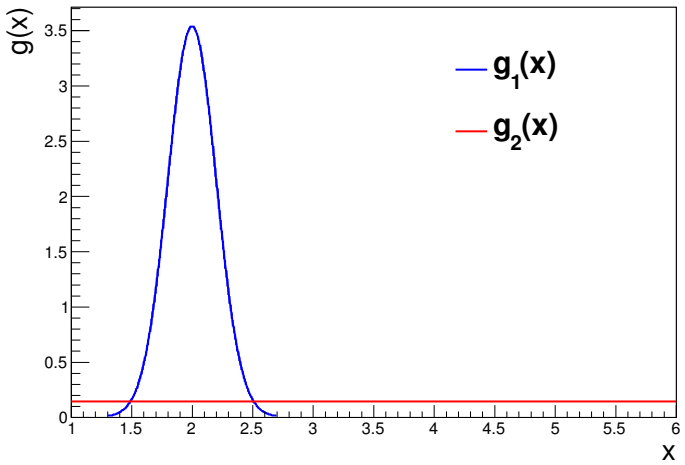


FIGURE 3.1: Probability density functions for signal and noise.

We calculate numerically the estimated fraction for three mock experiments in which 30, 300 and 3000 events[4] are drawn from a mixed distribution composed by EQUATION 3.28 (signal) and EQUATION 3.29 (noise). In FIGURE 3.2 we show the data in three typical runs in which the true signal to noise ratio is 0.8.

[4]The number of events in the experiments are selected to be close to the number of events in the $X_{\max}$ composition analysis with actual data of the Pierre Auger Observatory at energies around 1, 10 and 100 EeV.

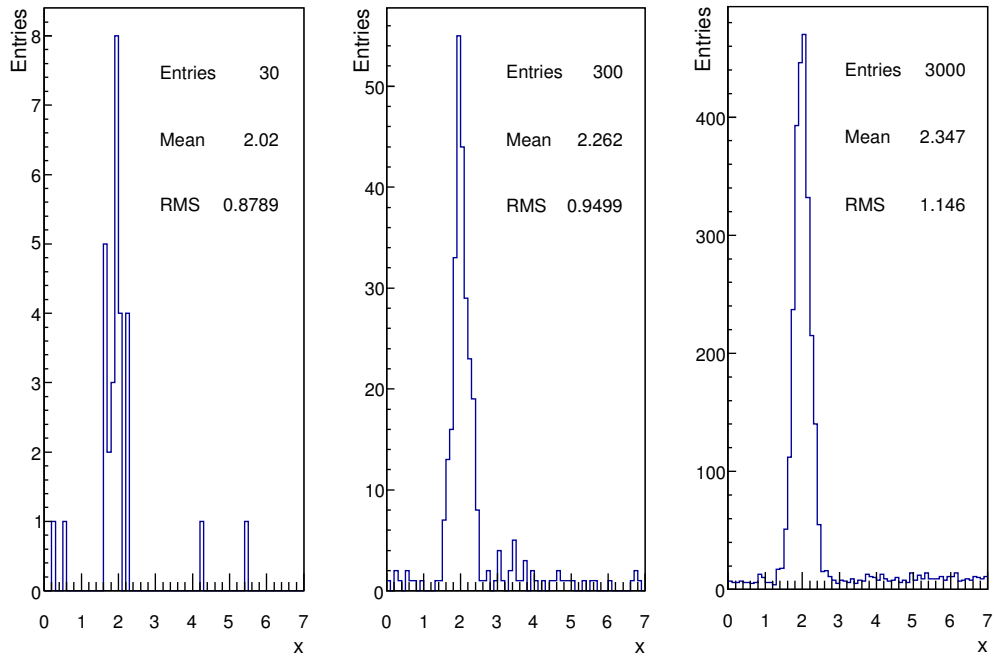


FIGURE 3.2: Data histograms for each of the three mock experiments each having 30, 300 and 3000 events sampled from the distributions in FIGURE 3.1 with a signal to noise ratio of 0.8.

In TABLE 3.1 we show the results of the inference of the different methods discussed above for the signal to noise ratio (α) for the three mock runs. Note that all estimators give a reasonable estimate of the true fraction, but the best are those based on the posterior p.d.f. Note that the χ^2 method is the worst estimator for this example. The bin size of 0.002 has been chosen using the χ^2 method, which is quite unreasonable for small number of events. This was done on purpose to show that one does not need to bin the data, and that binning can produce bad results, if poorly done.

# Events	$\langle\alpha\rangle$	$\alpha_{\max}$	α_{χ^2}	$\alpha_{<>}$
30	0.82	0.84	0.00	0.70
300	0.79	0.79	0.63	0.82
3000	0.80	0.80	0.81	0.78

TABLE 3.1: Results for the signal/noise discrimination.

In FIGURE 3.3 the posterior probability distributions for α are shown for the three mock runs and in FIGURE 3.4 the χ^2 functions of α are also displayed for comparison.

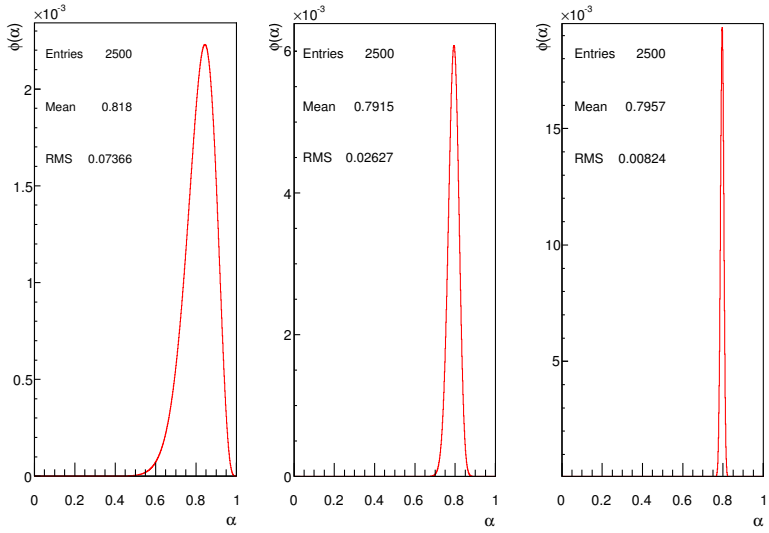


FIGURE 3.3: Probability functions of α for the three mock runs with 30, 300 and 3000 events. The true fraction is 0.8 in the three cases.

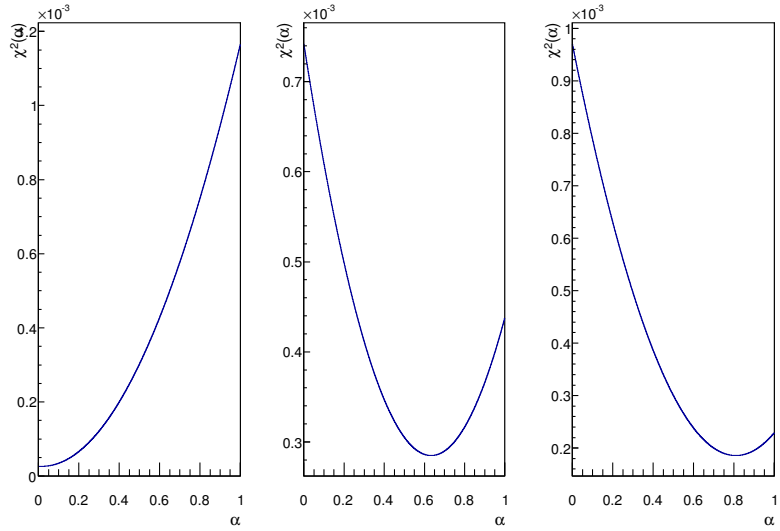


FIGURE 3.4: Same as FIGURE 3.3 but the χ^2 functions are shown instead of the posterior distributions.

3.3.2 Mixture of two signals and distance parameter

As a further example we now consider the problem of discrimination of two signals modelled as two Gaussian distributions of different means and widths

$$g_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}, \tag{3.30}$$

$$g_2(x) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}. \tag{3.31}$$

A number of "distance measures" for probability density functions has been proposed in the literature. We here define the distance between two distributions $g_1(x)$ and $g_2(x)$ as the non-overlapping area:

$$d_1(g_1, g_2) = \int dx|g_1(x) - g_2(x)|, \tag{3.32}$$

which ranges between 0 and 2. For $d_1 = 2$ the distributions do not overlap; for $d_1 = 0$ the distributions are identical. In the Appendix A we discuss some other possibilities and justify the choice of the overlapping area, as our distance. For the previous example of heads and tails, the distance is given by $d_1 = 2(1-\delta-\epsilon)$, which is the pre-factor appearing in Equation 3.25.

We proceed to study the discrimination power of the chosen methods as a function of the distance between the probability distributions. The values of the parameters μ_1, σ_1 and σ_2 are fixed to $\mu_1 = 0$, $\sigma_1 = 1$, and $\sigma_2 = 0.5$ while μ_2 is varied from -1.5 to 1.5. The variation of the distance is given by the variation of μ_2. Different values of the composition fraction will be also used. For Gaussian distributions the distance between the two functions can be written as

$$d_1(g_1, g_2) = \int dx|g_1(x) - g_2(x)| = I_1 + I_2 + I_3, \tag{3.33}$$

where the I_i are combinations of error functions[5]

$$I_1 = \frac{1}{2}|\mathrm{Erf}(\frac{x_{c1}-\mu_1}{\sqrt{2}\sigma_1}) - \mathrm{Erf}(\frac{x_{c1}-\mu_2}{\sqrt{2}\sigma_2})|, \tag{3.34}$$

$$I_2 = \frac{1}{2}|\mathrm{Erf}(\frac{x_{c2}-\mu_1}{\sqrt{2}\sigma_1}) - \mathrm{Erf}(\frac{x_{c1}-\mu_1}{\sqrt{2}\sigma_1}) - \mathrm{Erf}(\frac{x_{c2}-\mu_2}{\sqrt{2}\sigma_2}) + \mathrm{Erf}(\frac{x_{c1}-\mu_2}{\sqrt{2}\sigma_2})|, \tag{3.35}$$

[5]The error function is defined as: $\mathrm{Erf(x)} = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-t^2}dt$.

$$I_3 = \frac{1}{2}|\mathrm{Erf}(\frac{\mu_1 - x_{c2}}{\sqrt{2}\sigma_1}) - \mathrm{Erf}(\frac{\mu_2 - x_{c2}}{\sqrt{2}\sigma_2})|, \tag{3.36}$$

Here x_{c1} and x_{c2} are the two solutions to the equation

$$g_1(x) = g_2(x).$$

To study the uncertainty in relation to the distance parameter We have simulated 10000 runs with 30, 300, and 3000 events for each fixed value of the composition fraction, α_{true}, varying from 0 to 1 in steps of 0.1 (then, a total of $5.28 \cdot 10^6$). The estimated value of α for each case is given by the four estimators described before: the value of α that gives the same mean, $\alpha_{<>}$, the value which minimises the χ^2 function in binned data, α_χ, the value that maximises the posterior distribution, α_{max}, and the expected value of α given the posterior distribution, $\langle\alpha\rangle$. In FIGURE 3.5-FIGURE 3.7 $|\alpha - \alpha_{true}|$ (where α is the estimated fraction) is shown as a function of the distance for 30, 300, and 3000 events. Note that for a large number of events the mean value $< \alpha >$ and the mode of the posterior p.d.f. α_{max} converge but for small number of events or small distances the mean performs slightly better.

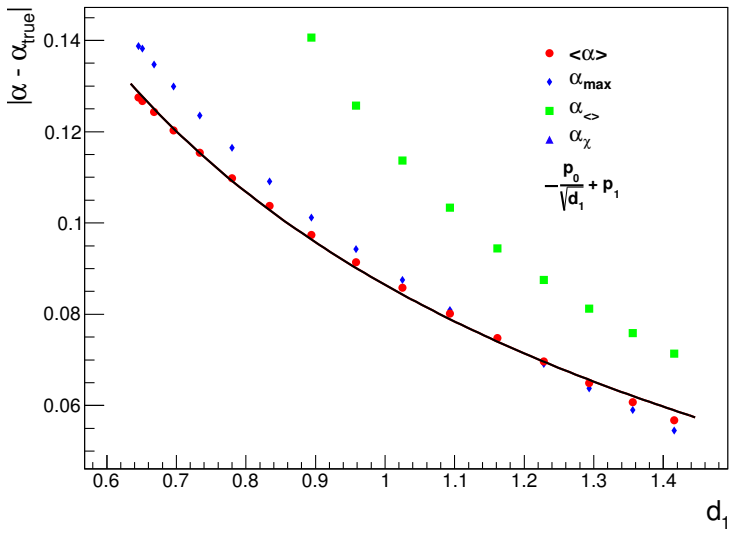


FIGURE 3.5: Absolute difference between the estimated fraction and the true fraction as a function of the distance between the distributions for 10000 runs each with 30 data sample. Note that $d_1 = 0$ means that the two distributions are equal while $d_1 = 2$ means that the distributions are completely separated. We remark that for the case using χ^2 minimisation is outside the range of the figure.

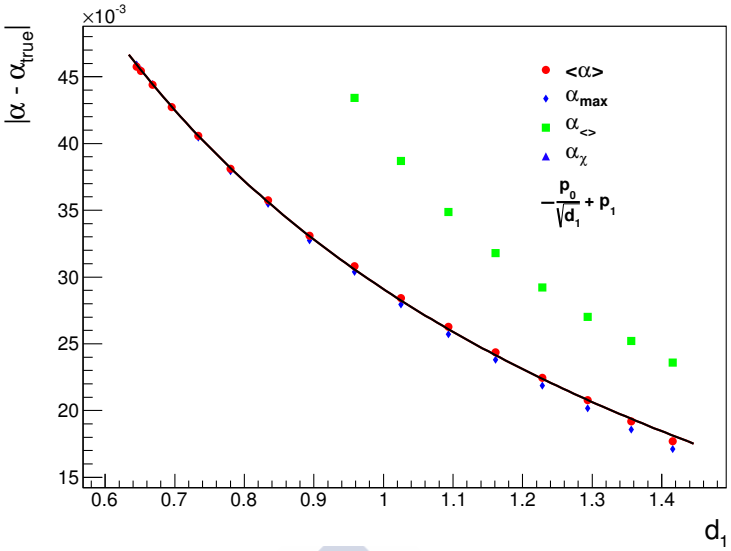


FIGURE 3.6: Same as FIGURE 3.5 but the data set contains 300 events at each run. Note again that the χ^2 minimisation is outside the range of the figure.

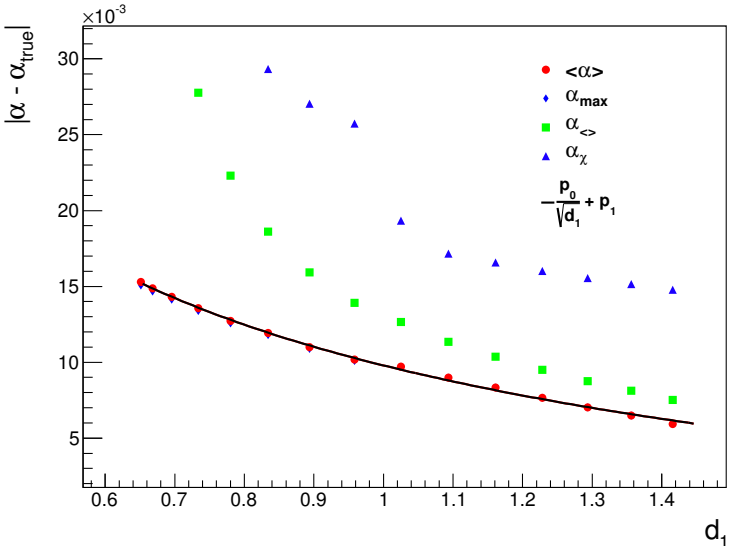


FIGURE 3.7: Same as FIGURE 3.5 but each run is analysed with 3000 events.

The χ^2 method does not show in FIGURE 3.5 nor in FIGURE 3.6, it is off-scale due to the artificially low binning used in the estimation of the χ^2. In FIGURE 3.7 it is clearly shown as the worst method. One can see in FIGURES 3.5-3.7 that both $\langle\alpha\rangle$ and α_{max} approximately scale as the square root of the distance. A fit to the function

$$|\alpha - \alpha_{true}| = \frac{p_0}{\sqrt{d_1}} + p_1, \tag{3.37}$$

is shown in all the figures. This is in agreement with the results of SECTION 3.2.1[6] and justifies our choice for the definition of distance. The uncertainty of $\alpha_{<>}$ more than doubles that of the maximum likelihood (α_{max}) or the mean value of the posterior p.d.f. ($\langle\alpha\rangle$) for small distances.

In a practical situation we cannot study $|\alpha - \alpha_{true}|$ and the standard deviation of the posterior probability functions can naturally give us an estimate of the uncertainty for a single trial. When we calculated the posterior distributions in the above examples for all the trials we also calculated their standard distributions. The averages of the corresponding standard deviations of the posterior probability distributions as a function of the distance are shown in figure 3.8.

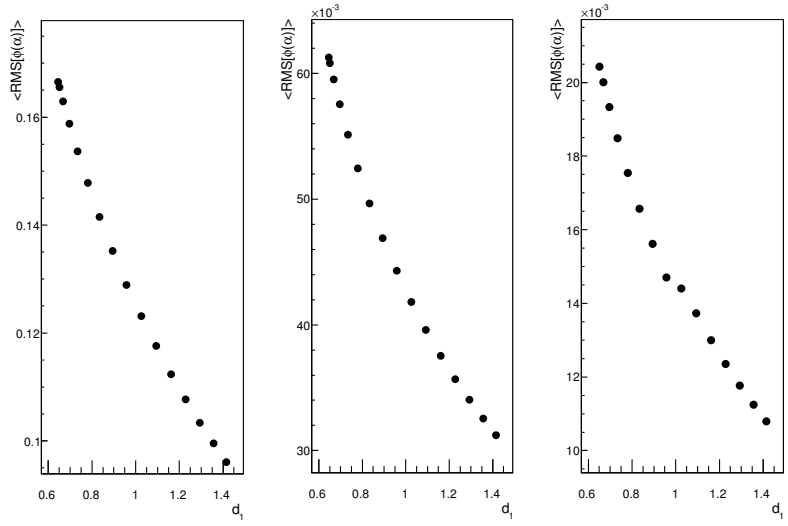


FIGURE 3.8: Averaged standard deviation of the posterior probability distribution for all trials as a function of the distance between the probability distributions. Left: analysis done with 30 events. Middle: 300 events. Right: 3000 events.

We now examine a single example corresponding to a mixture of two Gaussians with different mean and standard deviation. In this case, the values of the parameters chosen are $\mu_1 = 2$, $\sigma_1 = 0.2$, $\mu_2 = 2.3$ and $\sigma_2 = 0.4$. The distance between the distributions is $d_1 = 0.926$. In FIGURE 3.9 the probability density functions for $g_1(x)$ and $g_2(x)$ considered for this case are shown. Samples of data distributions are shown in FIGURE 3.10. In TABLE 3.2 we show the results for the samples shown in

[6]In SECTION 3.2.1 was shown that $\langle\alpha\rangle = \frac{1}{1-\delta-\epsilon}\left(\frac{B(1-\epsilon,n_1+2,n_2+1)-B(\delta,n_1+2,n_2+1)}{B(1-\epsilon,n_1+1,n_2+1)-B(\delta,n_1+1,n_2+1)}\right)$. In this case the distance between the two probability distributions is $2(1-\delta-\epsilon)$. As the distance increases the difference $\alpha_{true} - \langle\alpha\rangle$ becomes in a Gaussian with mean 0 and standard deviation proportional to $\frac{1}{\sqrt{1-\delta-\epsilon}}$

FIGURE 3.10. In FIGURE 3.11 and FIGURE 3.12 we respectively show the posterior probability distributions and the χ^2 distributions.

Accordingly to our previous discussion by looking at FIGURES 3.5-3.7, one expects the fraction to be estimated with an uncertainty of ~ 0.1, 0.03 and 0.01 respectively for 30, 300 and 3000 events when using the Bayesian estimators $\langle\alpha\rangle$ or α_{max}. For the $\alpha_{<>}$ estimator the uncertainty expected is < 0.14, 0.045 and 0.014.

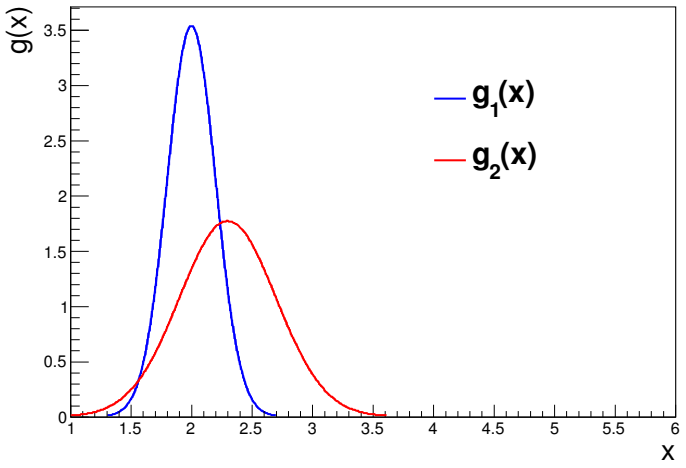


FIGURE 3.9: Probability density functions for the two Gaussians with parameters $\mu_1 = 2$, $\sigma_1 = 0.2$, $\mu_2 = 2.3$ and $\sigma_2 = 0.4$.

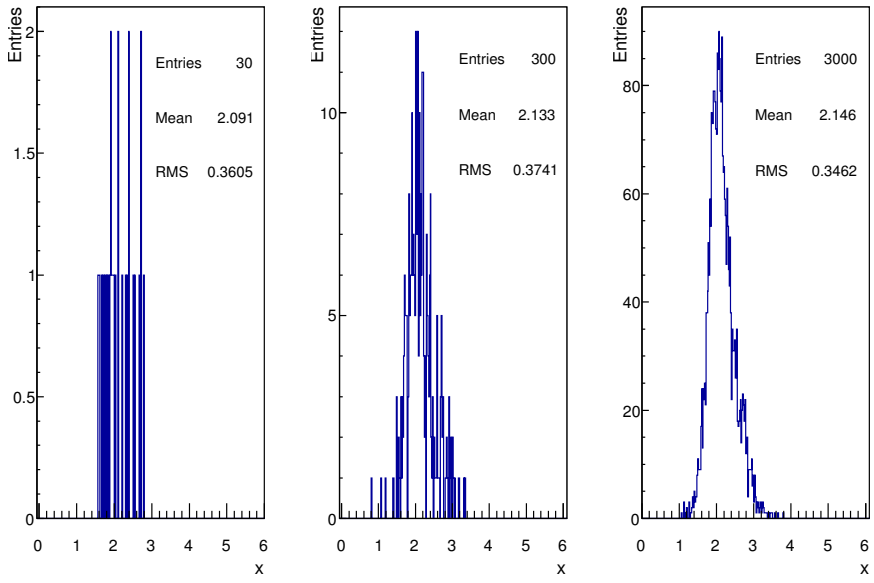


FIGURE 3.10: Data distributions analysed for 30, 300 and 3000 events corresponding to the probability distributions in FIGURE 3.9.

# Events	$\langle\alpha\rangle$	$\alpha_{\max}$	α_{χ^2}	$\alpha_{<>}$
30	0.46	0.48	0.0	0.70
300	0.51	0.52	0.50	0.55
3000	0.51	0.51	0.54	0.51

TABLE 3.2: Results applying the different estimations to infer α from the data samples. Here the true fraction is $\alpha_{true} = 0.5$.

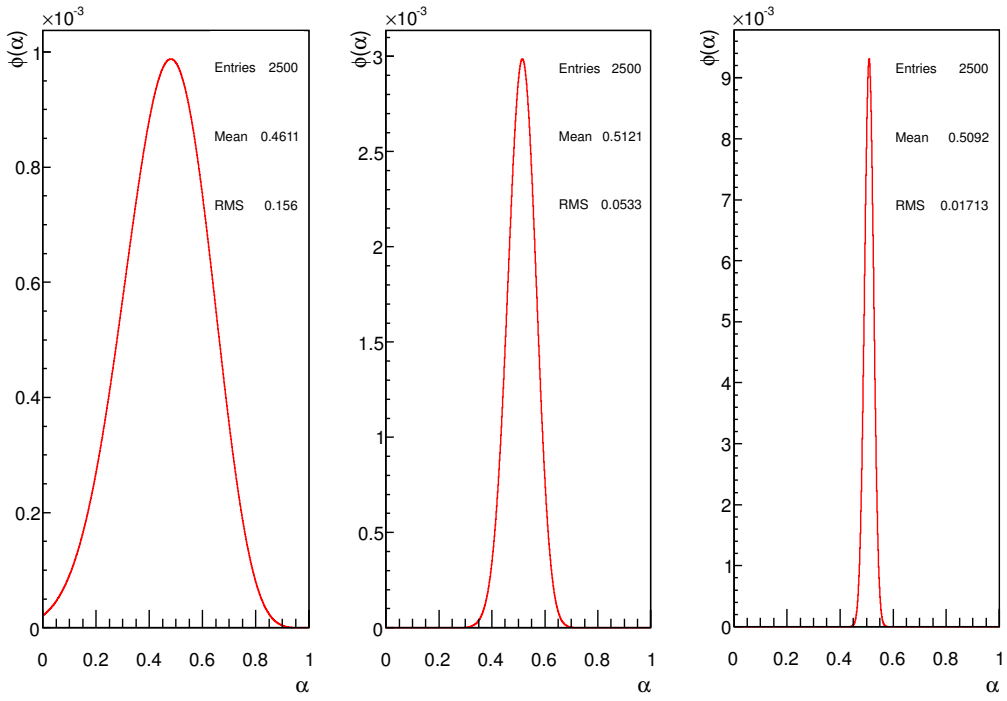


FIGURE 3.11: Posterior probability distributions of α for 30, 300, and 3000 events. The true fraction is $\alpha_{true} = 0.5$.

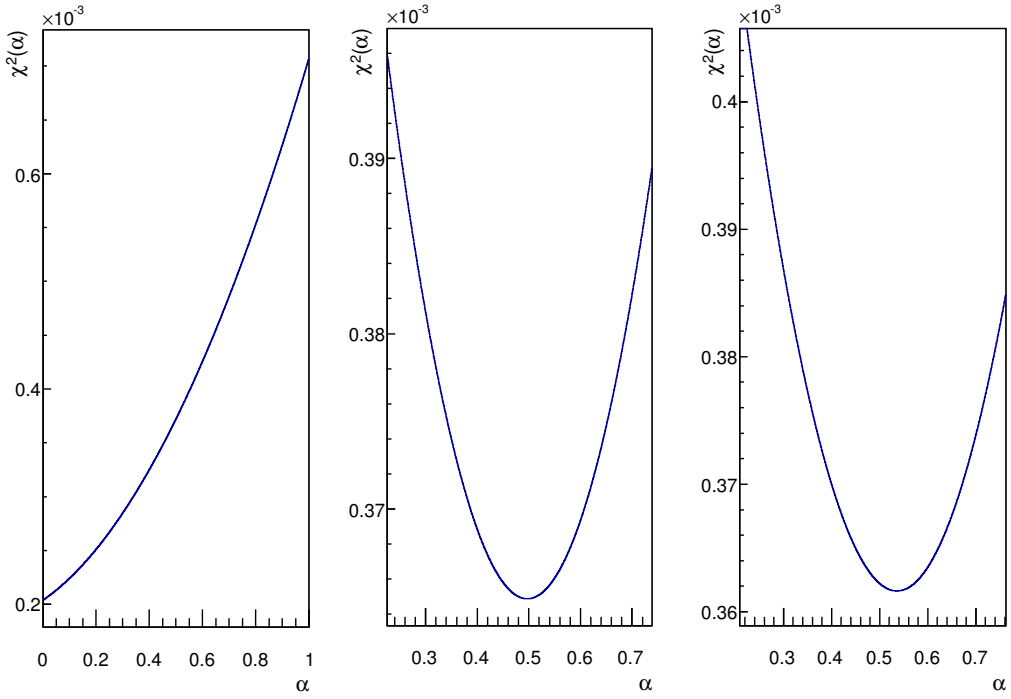


FIGURE 3.12: χ^2 functions for the same cases as FIGURE 3.11.

3.4 Analysis of composition using $X_{\max}$ distributions

In the previous sections we have shown that the best estimators for the fraction of a mixture of two components are both the mean and maximum values of the posterior p.d.f., $\pi(\alpha|D)$. We will now make an analysis of composition of the high energy cosmic rays using the maximum of the longitudinal development profile, $X_{\max}$ as our discriminator [59]. We have used all the estimators discussed previously but we will concentrate on the results using the mean and the maximum of the posterior p.d.f. The $X_{\max}$ distributions are generated with the CONEX [60] generator using EPOS LHC [49] as the hadronic model.

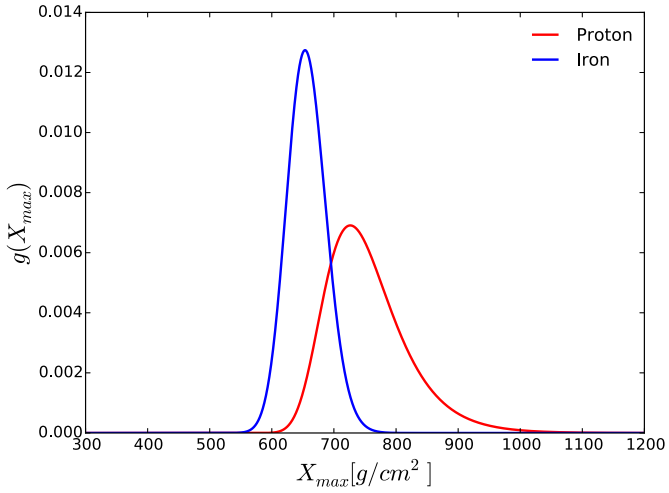


FIGURE 3.13: $X_{\max}$ probability density functions for proton (red) and iron (blue) using EPOS LHC hadronic interaction model at energies between $\log_{10}(E/\text{eV}) = 18$ and $\log_{10}(E/\text{eV}) = 18.1$.

Consider a typical example of a two-component mixture in which we want to find the proton and iron fractions in a data sample corresponding to energies between 1 EeV to 1.25 EeV. The two distributions of $X_{\max}$ are shown in FIGURE 3.13. They have been simulated using EPOS LHC model. The distance d_1 between the simulated distributions model in this energy bin is 1.55. Then we can use, as a rule of thumb, our estimated resolution in the composition fraction approximated by $|\alpha - \alpha_{\text{true}}| \sim 1/\sqrt{Nd_1}$ which amounts to 0.05, 0.012, and 0.006 respectively for 30, 300, or 3000 events (see FIGURES 3.5-3.7). The standard deviation of the posterior distributions is respectively expected to be of order $\sigma \sim 0.1$, 0.03, 0.01 (as shown in FIGURE 3.8).

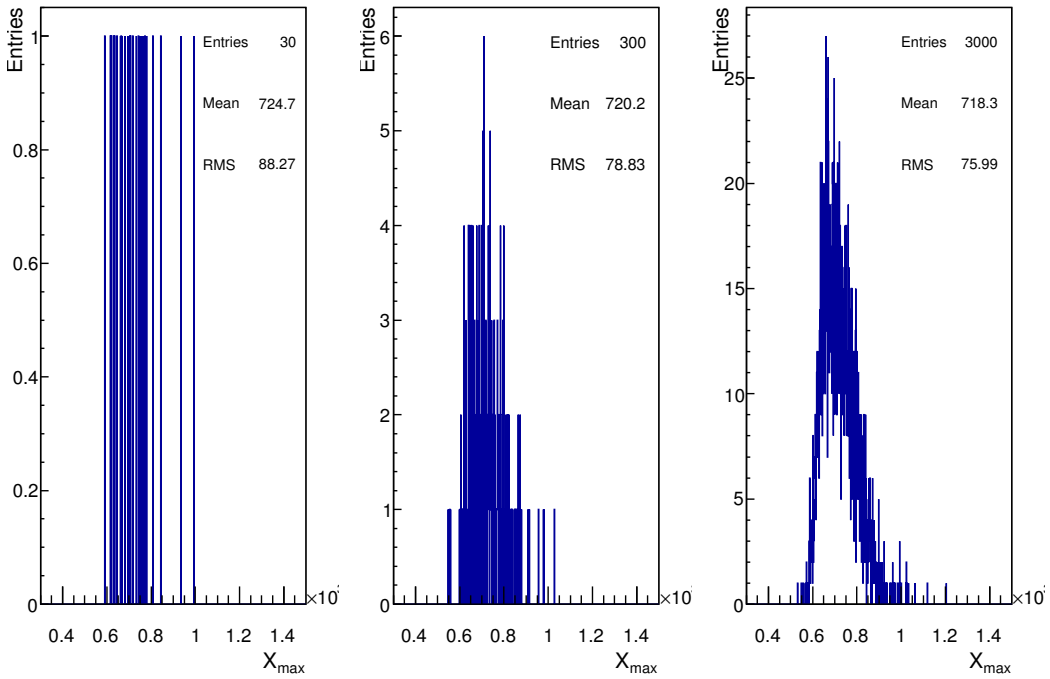


FIGURE 3.14: Sample data distributions analysed for 30, 300, and 3000 events corresponding to the distributions in Fig.3.13.

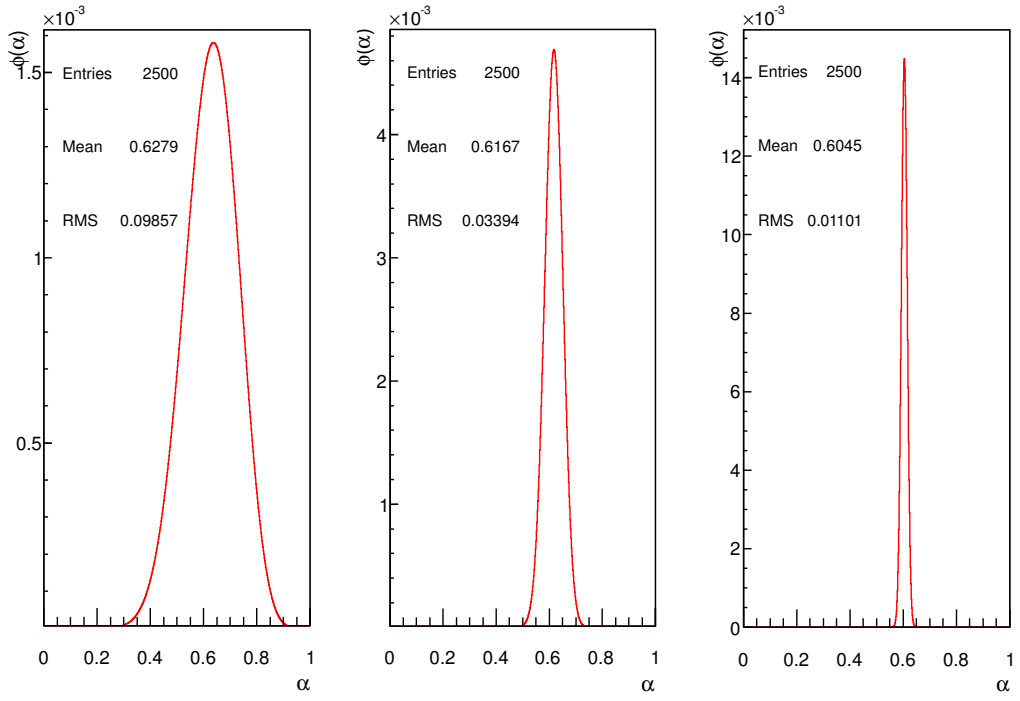


FIGURE 3.15: Posterior probability functions of α for 30, 300, and 3000 events. The number of entries in the plots corresponds to the number of selected points of α to evaluate the posterior probability function. In this case we have discretised the continuous space of α into 2500 points.

Data histograms and posterior probability density functions are respectively shown in FIGURE 3.14 and FIGURE 3.15. In TABLE 3.3, we also show our results for these analyses where the uncertainties are calculated in the following way: for the uncertainty in the mean value we take the standard deviation of the posterior distribution, for the uncertainty in the maximum likelihood value we take the width at 68% confidence level around the mode of the posterior distributions. In these examples, both

# Events	$\langle\alpha\rangle$	$\alpha_{\max}$
30	0.63 ± 0.10	0.64 ± 0.10
300	0.62 ± 0.03	0.62 ± 0.03
3000	0.604 ± 0.010	0.604 ± 0.010

TABLE 3.3: Composition fraction obtained with the mean and maximum values of the posterior probability function of α. In this case $\alpha_{true} = 0.6$.

the mean and the maximum of the posterior probability give the same results.

3.5 Study of methods with more than 2 primaries

In the previous sections different estimators for the evaluation of composition fraction of a mixture of two probability distributions have been compared. It has also been shown that the best estimators are the mean and the mode of the posterior probability density function. The χ^2 method gives results comparable to the maximum likelihood estimator (mode of the posterior p.d.f. when the prior $\pi(\alpha)$ is flat in α) if the number of events is large, but we have also seen that with an inadequate binning of data the results can be misleading. A remarkable results followers: with few events, the mean value of the probability distribution is the best estimator. A measure of the distance between the two probability distributions has been studied and has been shown how it can give us an estimation of the discrimination power for two distributions (see Equation 3.37). If the distance d_1 is small, the discrimination between the two compositions will be poor. If the distance is large it will be optimal. We have shown that as a "rule of thumb" the discrimination power scales as $1/\sqrt{d_1 N}$ with N the number of events.

In this section the discrimination power of the best estimators (maximum likelihood and mean of the posterior p.d.f.) are studied in the case of mixtures of three and four distributions corresponding to three and four different primaries.

To be concrete, the $X_{\max}$ distributions generated with Epos LHC will be used for each primary (see Figure 3.16). For this analysis two scenarios are considered. Firstly the mixture of three $X_{\max}$ distributions generated by hydrogen, helium and iron nuclei are considered. To study the discrimination power with four primaries the $X_{\max}$ distribution of nitrogen is added.

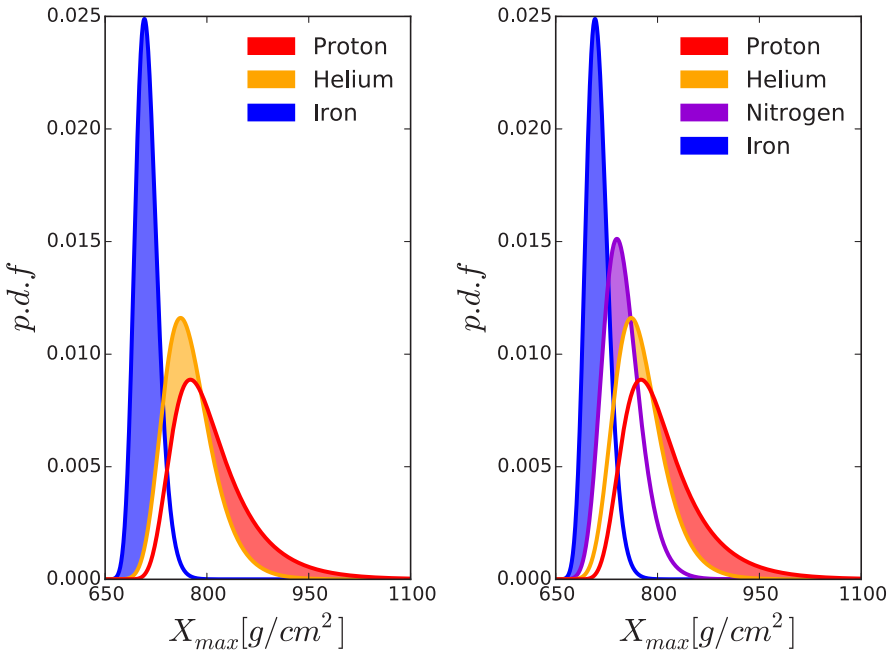


FIGURE 3.16: EPOS LHC $X_{\max}$ distributions for different primaries: proton (red), helium (orange), nitrogen (violet) and iron (blue). The shaded regions correspond with the not overlapped areas of the different distributions. The left panel corresponds with a scenario where only proton, helium and iron are considered. In the right panel nitrogen has been added.

It is straightforward to extend EQUATION 3.32 to more than two primaries. The area that does not overlap for a primary "p" is calculated as:

$$d_1^{(p)} = \int |g_p(x) - \max(\{g_i(x)\})| dx \tag{3.38}$$

where $g_i(x)$ is the set of probability distributions of the primaries different than "p". In an scenario we can calculate the area that does not overlap with any other primary and see that it is largest for the iron distribution followed by proton as it has been illustrated by the shaded areas in FIGURE 3.16. The resolution of the deduced fraction can be expected to follow the same ordering. In the four-primary-scenario the area that does not overlap for the nitrogen distribution is larger than that of helium, then the resolution is expected to be better for the nitrogen than for the helium.

As in SECTION 3.3 the absolute difference between the estimated fraction and the actual fraction of the simulated data is used to calculate statistically the resolution of the methods by making multiple simulations. In FIGURE 3.17 the average value of the resolution for the different primaries is shown as a function of the number of events

of the simulated data sample. The average is taken over 5000 trials with random fractions uniformly distributed over the 2-simplex and the 3-simplex [7]. One can see that again the best estimator is the mean of the posterior p.d.f. The difference between the resolution using the maximum or the mean also increases when the number of events decreases.

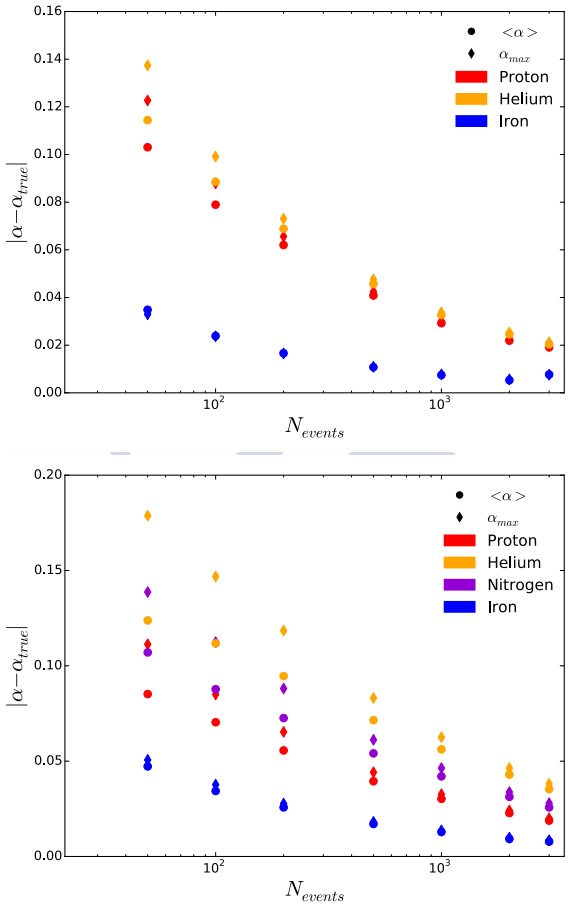


FIGURE 3.17: Average value of the absolute difference between the estimated fraction and the actual fraction as a function of the number of events in data using as estimators the maximum (diamonds) and the mean (circles) of the posterior p.d.f for the two considered scenarios; upper panel: proton in red, helium in orange and iron in blue; lower panel: proton in red, helium in orange, nitrogen in violet and iron in blue.

[7] A n-simplex is the generalisation of a n-dimensional triangle with $n+1$ vertices. The available volume of parameter space is not trivial because the fractions of all primaries must add to one. For two primaries it is the line $x + y = 1$. For three primaries it is the a triangle defined by the plane $x+y+z = 1$ and limited to $0 \leq x \leq 1$, $0 \leq y \leq 1$ and $0 \leq z \leq 1$. See APPENDIX B for more details.

$$\mathcal{S}^n = \{\vec{x} = (x_1, ..., x_{n+1}) \in \mathbb{R}^{n+1} \mid x_i \geq 0;\ \sum_{i=1}^{n+1} x_i = 1\}$$

3.6 The determination of confidence intervals in the Bayesian and Frequentist approaches

The list of references on confidence intervals is large, see, for instance, [61] and [62]. For completitude, a quick review is shown here. Let the data distribution be composed by two different primaries (distributions) and then the posterior p.d.f of the composition fraction given the data x is

$$P(\alpha|x) = \frac{P(x|\alpha)P(\alpha)}{P(x)}. \tag{3.39}$$

The problem of interest now is to get a confidence interval $[\alpha_{low}, \alpha_{up}]$ as small as possible where the true value of α is contained with a given probability q. Mathematically it is expressed as

$$P(\alpha \in [\alpha_{low}, \alpha_{up}]) = q. \tag{3.40}$$

The main differences between the Bayesian and the Frequentist procedure is how to calculate EQUATION 3.40. If we have made a measurement x, the probability density function that describes the data distribution is $P(x|\alpha)$ (remember that α is the unknown parameter).

From a Bayesian point of view, once the posterior probability density function is calculated, the confidence interval is given by

$$\int_{\alpha_{low}}^{\alpha_{up}} P(\alpha|x)d\alpha = q. \tag{3.41}$$

From the Frequentist point of view, the pdf in α does not exist. Any "prior" information about α is regarded as "subjective" information and should not be considered. Instead, the confidence interval is constructed using information on possible measurements.

$$\int_{x_{low}}^{x_{up}} P(x|\alpha)dx = q \tag{3.42}$$

It must be interpreted as

$$P(x \in [x_{low}, x_{up}]|\alpha) = q, \tag{3.43}$$

i.e, the probability of measuring x belonging to the interval $[x_{low}, x_{up}]$ is q, when the true value of the parameter is α. Note that the interpretations and the procedures are very different and, as it is expected, the limits must be different in each case. In

Section 3.6.2 we discuss under what circumstances the limits provided by the two methods can be equal.

3.6.1 Fitting the mass composition fraction

To simplify the problem a two composition scenario is assumed like in Section 3.3.2. Suppose that the cosmic rays arriving to Earth from the space can be of only two types: type "1" (iron) or of type "2" (protons) denoting by $g_1(x)$ the probability distribution of some measurable variable for the particles of type "1" and by $g_2(x)$ for the particles of type "2". This variable could, for instance, be X_{max} or N_μ which are currently used for estimating the composition in the Pierre Auger Observatory. The normalised probability distribution function is

$$P(x|\alpha) = \alpha\, g_1(x) + (1-\alpha)\, g_2(x). \tag{3.44}$$

where α is the composition fraction, the fraction of iron events. We have chosen for g_1 and g_2 two Gaussian distributions one with mean $\mu_1 = 0$ and variance $\sigma_1^2 = 1$ and the other with mean $\mu_2 = 1$ and variance $\sigma_2^2 = 2$ which allow analytic results for most cases. In Figure 3.18, the $g_i(x)$ chosen for our example are explicitly shown.

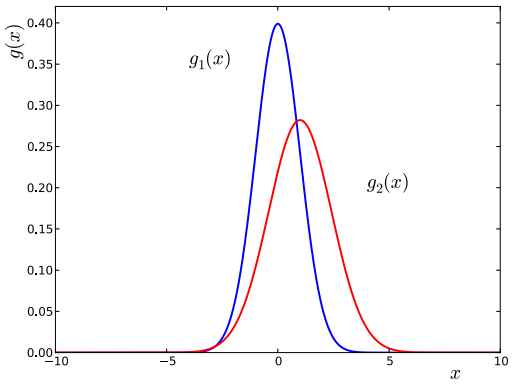


Figure 3.18: Probability density functions. $g_1(x)$ is blue and $g_2(x)$ is red (see text).

Let us now consider the confidence intervals obtained with the two methods discussed above after a single measurement of the observable has been done. The lower limit

for α at q confidence level must be understood as the value $\alpha = \alpha_0$ such that the probability of obtaining α larger than α_0 is q. In the Bayes case a flat prior is assumed, which is equivalent to giving the same weights to all possible composition fractions before the measurements are made. The Bayesian lower limit is given by α_0 satisfying

$$\int_{\alpha_0}^{1} P(\alpha|x_0)d\alpha = q, \tag{3.45}$$

where x_0 is the measured variable. Solving EQUATION 3.45, the α_0 obtained is

$$\alpha_0 = \frac{-g_2(x_0) + \sqrt{g_1(x_0)^2 - q[g_1^2(x_0) - g_2^2(x_0)]}}{g_1(x_0) - g_2(x_0)} \tag{3.46}$$

In the Frequentist approach the lower limit (Neyman's limit) is given by value of α satisfying

$$\int_{x_0}^{\infty} P(x|\alpha_0)dx = q \tag{3.47}$$

Solving the above expression for α_0 one can obtain the curve which defines the lower limit

$$\alpha_0 = \frac{q - 1 + G_2(x_0)}{G_2(x_0) - G_1(x_0)}, \tag{3.48}$$

where $G_i(x_0) = \int_{-\infty}^{x_0} g_i(x)dx$ is the cumulative distribution of $g_i(x)$. By looking at EQUATION 3.46 and EQUATION 3.48 it should be clear that the expressions for the lower limit α_0 (as a function of the measured value, x_0) are very different in the Bayesian and in the Frequentist point of views. One cannot expect to get the same limits. In FIGURE 3.19, the limits calculated with EQUATION 3.46 and EQUATION 3.48 at $q = 0.5$ and $q = 0.9$ confidence level are shown in the two dimensional plot of x_0 and α_0. As it is apparent in the figure, they are completely different.

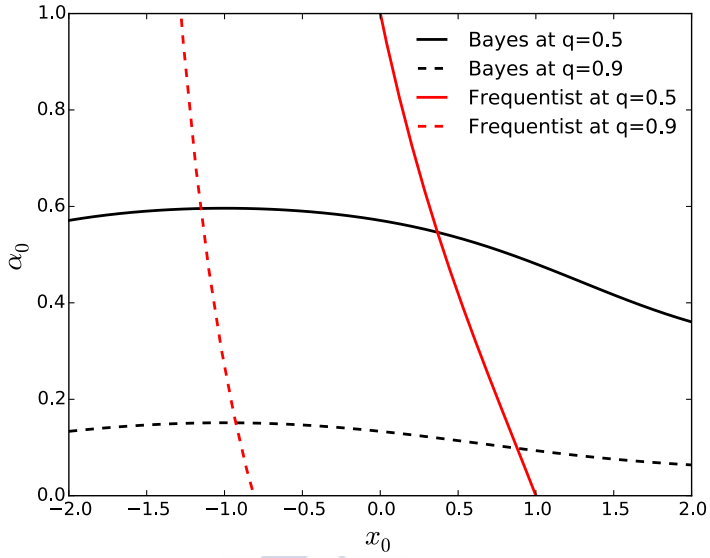


FIGURE 3.19: Lower limits using a Bayesian approach (black) and Frequentist approach (red) for 50% (continuous lines) and for 90% (dashed lines) of confidence levels.

In FIGURE 3.19 only the physical region is shown. It is well known that the Neyman construction of confidence levels can give unphysical values. One can see that EQUATION 3.48 is not limited to the physical region where $\alpha \in [0, 1]$. In fact, notice that if $G_1(x_0) = G_2(x_0)$ then the limit on α grows without bound. On the other hand EQUATION 3.46 always gives results in the physical region, by construction. It appears that when $g_1(x_0) = g_2(x_0)$ there is a singularity, but in fact it is not the case. If $g_1(x_0) = g_2(x_0)$ the posterior probability density function is $P(\alpha|x_0) = 1$ if $\alpha \in [0., 1.]$ and 0 otherwise. That means that the posterior is just the prior and the datum does not modifies our knowledge about the parameter, obtaining as lower level

$$\alpha_0 = 1 - q \tag{3.49}$$

In [61] the authors propose a new construction of the region of integration in x based on an ordering parameter

$$R = \frac{\mathcal{L}(\alpha|x)}{\mathcal{L}(\alpha_{best}|x)} \tag{3.50}$$

where $\mathcal{L}(\alpha|x) = P(x|\alpha)$ is the likelihood function and α_{best} is the composition fraction which maximises the likelihood function, nevertheless these method does not

guarantee the exclusion of the non-physical region unless this region was exclude by hand (which is not different to take a prior). In fact one can check by using the ROOT [57] package that the Feldman-Cousins procedure does not hep to solve this problem, resulting in either the full parameter space or in an empty set.

Suppose that the measurement is $x_0 = 0.1$ and that one wants to obtain a 50 % confidence limit, then in view of FIGURE 3.19, the Frequentist limit is well within the physical region (and the procedure proposed in [61] is not expected to modify the results). Then one can read in FIGURE 3.19 that $P(\alpha \in [0.85, 1]) = P(\alpha \in [0, 0.85]) = 0.5$. This is a surprising result. At 50% C.L. it is possible to assure that $\alpha > 0.85$ by measuring a single event! If the single measurement was $x_0 = -1.1$ one could be ruling out more than half of the range at 90% C.L.! Notice also the slope of the curves. By changing the measured value from -1.1 to -1.3, the range of allowed values at 90% of C.L. is changed from $\alpha > 0.6$ to $\alpha > 0.8$. Such a high slope can not represent the true information we get from the measurement of a single event. However, consider the Bayes result. At 50% C.L. with $x_0 = 0.1$, $P(\alpha \in [0.56, 1]) = P(\alpha \in [0, 0.56] = 0.5$ is found, *i.e.* it is a little bit more probable that the event measured is iron and not proton. Notice than in this case the measurement is less than one sigma away from the proton distribution. The Bayes approach gives at 90% C.L. that the fraction of iron events is greater than 0.15 while in the Frequentist approach the fraction of events is greater than 0.75 at the same C.L., which is again difficult to accept.

Consider now some scenario in the case when two events are measured, x_1, x_2. In the Bayesian approach there is no major difficulty in setting confidence limits. Assuming that the two events are independent, then the probability of obtaining the two values is

$$\begin{aligned} P(x_1, x_2|\alpha) &= (\alpha\, g_1(x_1) + (1-\alpha)\, g_2(x_1)) \\ &\times (\alpha\, g_1(x_2) + (1-\alpha)\, g_2(x_2)), \end{aligned} \tag{3.51}$$

and the limits are calculated in the same way. In the Frequentist approach, it is not clear how to proceed. The standard way would be to evaluate the mean value $y_2 = (x_1 + x_2)/2$ and set the limits using this variable solely. This can be done in both the Frequentist and the Bayesian approaches, but notice that in the example the mean is not a sufficient statistics and therefore it does not include all the information available of the problem. For instance, it is clear that getting $x_1 = x_2 = 1$ should give rather different limits than $x_1 = 0$, and $x_2 = 2$, despite having the same mean. Since

in our example the two distributions are Gaussians, the distribution of the mean is also a combination of Gaussians giving

$$\begin{aligned} P(y_2|\alpha) &= \int\int P(x_1, x_2|\alpha)\delta(y_2 - \frac{x_1 + x_2}{2})dx_1 dx_2 \\ &= \alpha^2 f_1(y_2) + 2\alpha(1-\alpha)f_2(y_2) + (1-\alpha)^2 f_3(y_2), \end{aligned} \tag{3.52}$$

where the f_i are Gaussian distributions. f_1 has mean μ_1 and $\sigma^2 = \sigma_1^2/2$, f_3 has mean μ_2 and $\sigma^2 = \sigma_2^2/2$, and f_2 has mean $(\mu_1 + \mu_2)/2$ and $\sigma^2 = (\sigma_1^2 + \sigma_2^2)/4$.

In FIGURE 3.20 the 50 % and 90 % C.L. for both the Bayesian and the Frequentist approach are again shown. Once more one can see that in the frequentist approach the confidence interval of α has a high slope and gives non-physical values for most of the y_2 range. The Bayesian results however, give physical results for any value of y_2 within the available range.

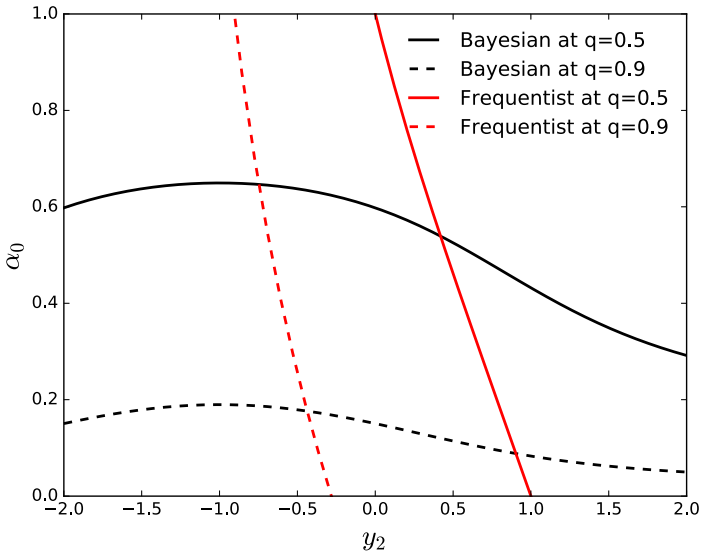


FIGURE 3.20: Lower limits using a Bayesian approach (black) and Frequentist approach (red) for 50% (continuous lines) and for 90% (dashed lines) of confidence levels as a function of the mean of two events.

In the case of n large measured events, the central limit theorem guarantees that the mean value $y_n = 1/n \sum x_i$ will have a Gaussian distribution of mean

$$\bar{y}(\alpha) = \alpha\mu_1 + (1-\alpha)\mu_2 \tag{3.53}$$

and sigma

$$\sigma_n^2 = \frac{\sigma_0^2}{n} = \frac{1}{n}\left[\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \alpha(1-\alpha)(\mu_1-\mu_2)^2\right]. \tag{3.54}$$

where μ_i, σ_i are the mean and sigma of the original distributions $g_i(x)$. Repeating the previous calculations the limits for the Frequentist approach are given implicitly by

$$\bar{y}(\alpha) = y_n \pm \sigma_n H \tag{3.55}$$

where y_n is the measured value of the mean of n events and H is given by

$$H = \sqrt{2}\,\mathrm{Erfc}^{-1}(2q), \tag{3.56}$$

which is the inverse of the complementary error function at q confidence level. The meaning of the expression is rather clear. The limit at 68% confidence level is 1 σ above the mean and so on.

For the Bayesian approach the dependence of σ_n on α can be neglected. This is a good approximation since σ_n is small for large n. Then the limit is given by the implicit equation

$$\frac{\mathrm{Erf}(a - b\,\alpha) - \mathrm{Erf}(a-b)}{\mathrm{Erf}(a) - \mathrm{Erf}(a-b)} = q, \tag{3.57}$$

where $a = (y_n - \mu_2)/(\sqrt{2}\sigma_n)$ and $b = (\mu_1 - \mu_2)/(\sqrt{2}\sigma_n)$. In fig. 3.21 the results for the two approaches are shown. Only in this limiting case the Frequentist approach coincides with the Bayesian limit and gives a reasonable result in the physical region.

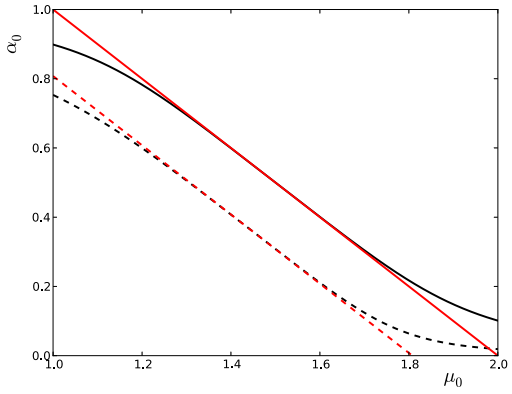


FIGURE 3.21: Lower limits using a Bayesian approach (black) and Frequentist approach (red) for 50% (continuous lines) and for 90% (dashed lines) of confidence levels as a function of the mean of n events.

3.6.2 Comments

For the determination of confidence intervals in the composition problem, the Frequentist approach can give results without physical meaning. Only in the limit of a large number of events, where it coincides with the Bayesian limit, the Frequentist approach gives reasonable results. A lot of tests and examples in the literature have shown the validity of the Frequentist approach, however one should notice that in most cases the quoted examples are such that integration over the "x" variable can be changed into an integration over the unknown parameter. This is true for the Gaussian case where integration over x can be trivially transformed into an integration over the mean of x and in other cases. In such cases the Bayesian and the Frequentist approach coincide (at least in the physical region of the parameters), and sensible results are expected. One could ask, however, why, if a single variable $x = x_0$ is measured, the confidence limit should depend on the probability of events that have never been measured. Apparently, this is the problem in the case at hand, integration over x has nothing to do with the integration over the parameter space α and one should expect Frequentist limits which are simply absurd.

Although this section has been concentrated on the estimation of the composition fraction, this same problem occurs, for instance, in the problem of signal-background discrimination or any other situation where the unknown parameter can not be cast into a change of the measured variables.

3.7 Extending the procedure to realistic detectors

The ideal detector does not exist. In general, when a detector measures some variable X, its estimate x_{obs} is different than the true value x_{true} because of the combination of various effects which are inherent to the measurement process. In a given experiment this applies to the relevant distribution $f(x_{true})$ which is related with the measured distribution $g(x_{obs})$ through

$$g(x_{obs}) = \int_{-\infty}^{\infty} \mathcal{M}(x_{obs}|x_{true}) f(x_{true}) dx_{true} \tag{3.58}$$

where $\mathcal{M}(x_{obs}|x_{true})$ is called the response function and sometimes is also called migration function. The response function describes the detector response to an observed event x_{obs} including the instrumental resolution and efficiency. Then, the response

function gives the probability of measuring x_{obs} if the actual value is x_{true}. Separating the response function into the efficiency function $\epsilon(x_{true})$ and the resolution function $\mathcal{R}(x_{obs}|x_{true})$ EQUATION 3.58 can be re-written as

$$g(x_{obs}) = \int_{-\infty}^{\infty} \epsilon(x_{true})\mathcal{R}(x_{obs}|x_{true})f(x_{true})dx_{true} \tag{3.59}$$

To infer the composition and, in general, any parameter of interest, it is necessary to take into account the effects of the detector over the data. In the following we explain how to take the detector into account in a Bayesian approach. First of all, we assume an ideal detector with maximal accuracy and efficiency, that is, an efficiency and response functions given by:

$$\epsilon(x) = 1;\ \forall x \tag{3.60}$$

$$\mathcal{R}(x_{obs}|x_{true}) = \delta(x_{obs} - x_{true}) \tag{3.61}$$

Then, the likelihood function is given by

$$P(x|\alpha) = \sum_{j}^{P} \alpha_j g_j(x), \tag{3.62}$$

where α is a vector in $P-$space and α_j is the fraction of events of the primary j whose theoretical distribution is $g_j(x)$ and $x = x^{true} = x^{obs}$. It is obvious that this scenario is the best that one can find to infer any parameter of interest. Then, the probability of the composition fraction given the data sample $D = \{x_i\}_{i=1}^{N}$ is

$$\begin{aligned} P(\alpha|D,I) &= \frac{\prod_{i=1}^{N} P(x_i|\alpha)P(\alpha|I)}{\int_A \prod_{i=1}^{N} P(x_i|\alpha)P(\alpha|I)d\alpha_1...d\alpha_P} = \\ &= \frac{\prod_{i=1}^{N} \sum_{j=1}^{P} \alpha_j g_j(x_i)P(\alpha|I)}{\int_A \prod_{i=1}^{N} \sum_{j=1}^{P} \alpha_j g_j(x_i)P(\alpha|I)d\alpha_1...d\alpha_P}, \end{aligned} \tag{3.63}$$

where $P(\alpha|I)$ is the prior distribution of α and A is its space.

Now we assume that the detector is not perfect adding an efficiency function. Then, the likelihood function is changed and must be normalised correctly over the x variable. The likelihood becomes

$$P(x|\alpha) = \frac{\sum_j^P \alpha_j g_j(x)\epsilon(x)}{\sum_j^P \alpha_j \int_{-\infty}^{\infty} g_j(x)\epsilon(x)dx}. \tag{3.64}$$

The posterior is given applying as usual the Bayes' theorem "posterior $\propto$ likelihood $\times$ prior"

$$P(\alpha|D,I) = \frac{\prod_{i=1}^{N}\sum_{j=1}^{P}\alpha_j g_j(x_i)\epsilon(x_i)P(\alpha|I)}{\int_A \prod_{i=1}^{N}\sum_{j=1}^{P}\alpha_j \int_{-\infty}^{\infty} g_j(x_i)\epsilon(x_i)P(\alpha|I)dxd\alpha_1...d\alpha_P}. \tag{3.65}$$

Finally a finite detector resolution is added. The observed data sample $D^{obs} = \{x_i^{obs}\}_{i=1}^{N}$ is different of the actual data sample $D = \{x_i\}_{i=1}^{N}$ due to the detector resolution. The likelihood and posterior distributions change now to

$$P(x^{obs}|\alpha) = \frac{\sum_j^P \alpha_j \int_{-\infty}^{\infty}\mathcal{R}(x^{obs}|x)g_j(x)\epsilon(x)dx}{\sum_j^P \alpha_j \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathcal{R}(x^{obs}|x)g_j(x)\epsilon(x)dxdx^{obs}} \tag{3.66}$$

$$= \frac{\sum_j^P \alpha_j \int_{-\infty}^{\infty}\mathcal{R}(x^{obs}|x)g_j(x)\epsilon(x)dx}{\sum_j^P \alpha_j \int_{-\infty}^{\infty} g_j(x)\epsilon(x)dx}, \tag{3.67}$$

$$P(\alpha|D^{obs},I) = \frac{\prod_{i=1}^{N}\sum_{j=1}^{P}\alpha_j \int_{-\infty}^{\infty}\mathcal{R}(x_i^{obs}|x)g_j(x)\epsilon(x)dxP(\alpha|I)}{\int_A \prod_{i=1}^{N}\sum_{j=1}^{P}\alpha_j \int_{-\infty}^{\infty} g_j(x)\epsilon(x)dxP(\alpha|I)d\alpha_1...d\alpha_P}. \tag{3.68}$$

Where it has been assumed that $\int_{-\infty}^{\infty}\mathcal{R}(x^{obs}|x)dx^{obs} = 1$ to obtain EQUATION 3.67 from EQUATION 3.66.

Up to now only the efficiency and detector resolution have been taken into account. The Bayesian approach allows us to explore more effects, for instance, we can assume that it is necessary to reject events outside some region [8] $\mathcal{C}$ for any reason after the data was collected. This cut can be described by a *characteristic* function of C

$$\chi_{\mathcal{C}}(x^{obs}) = \begin{cases} 1 & \text{if} \quad x^{obs} \in \mathcal{C} \\ 0 & \text{if} \quad x^{obs} \notin \mathcal{C} \end{cases} \tag{3.69}$$

The likelihood and the posterior distributions are now

$$P(x^{obs}|\alpha) = \frac{\sum_{j=1}^{P}\alpha_j\chi_{\mathcal{C}}(x^{obs})\int_{-\infty}^{\infty}\mathcal{R}(x^{obs}|x)g_j(x)\epsilon(x)dx}{\sum_{j=1}^{P}\alpha_j \int_{\mathcal{C}}\int_{-\infty}^{\infty}\mathcal{R}(x^{obs}|x)g_j(x)\epsilon(x)dxdx^{obs}} \tag{3.70}$$

$$P(\alpha|D^{obs},I) = \frac{\prod_{i=1}^{N}\sum_{j=1}^{P}\alpha_j\chi_{\mathcal{C}}(x^{obs})\int_{-\infty}^{\infty}\mathcal{R}(x_i^{obs}|x)g_j(x)\epsilon(x)dxP(\alpha|I)}{\int_A \prod_{i=1}^{N}\sum_{j=1}^{P}\alpha_j \int_{\mathcal{C}}\int_{-\infty}^{\infty}\mathcal{R}(x^{obs}|x)g_j(x)\epsilon(x)dxdx^{obs}P(\alpha|I)d\alpha_1...d\alpha_P}. \tag{3.71}$$

[8] The region $\mathcal{C}$ is used to describe any condition imposed on data, in general, if P is the condition, then $\mathcal{C}$ is defined as $\mathcal{C} = \{x^{obs}/P\}$. This implies that $\int \chi_{\mathcal{C}}(x_{obs})dx_{obs} = \int_{\mathcal{C}} dx_{obs}$ where the integral is limited to the $\mathcal{C}$ domain.

Two applications are going to be discussed. In the first example we are going to study how our inference is changing as our detector gets modified studying a single data set from a $X_{\max}$ mock distribution which is a mixture of two primaries. The composition fraction will be estimated using the mean of the posterior probability density function in successive steps as the different detector effects are progressively added to the detector or, equivalently, as the quality of data is impoverished.

In the second application several number of trials are performed in a four-composition scenario. Two approaches are compared. One requires using "anti-bias" cuts to eliminate the bias in the data in order to infer an unbiased composition. This is the current method used by the Auger Collaboration to deduce the composition of the $X_{\max}$ measurements and implies using a restricted data set minimising the selection bias. The second approach uses all the data set which has a selection bias but this is accounted for using the Bayesian analysis.

3.7.1 Study of detector effects on the composition estimation: step by step

The goal of this section is to compare the inference of the composition of a data sample as the quality of the data becomes worse by performing analyses in four scenarios. The mock data sample is initially composed of 1000 $X_{\max}$ measurements of showers generated by 80% protons and 20% iron nuclei. In each scenario the data sample recorded is modified by the different resolution and efficiency assumed for the detector as shown in FIGURE 3.24.

- Scenario A: an ideal or " perfect" detector is first assumed in the sense that its field of view is infinite, all the events are detected (the efficiency is always 1) and it has a perfect resolution. The data distribution coincides with a random sample of the actual distribution because the detector does not introduce any distortion on the data. To infer the composition only the theoretical $X_{\max}$ distributions for proton and iron are needed (denoted by $g_p(X_{\max})$ and $g_{Fe}(X_{\max})$). The likelihood function is

$$P(X_{\max}|\alpha) = \alpha g_p(X_{\max}) + (1-\alpha) g_{Fe}(X_{\max}), \tag{3.72}$$

  which is just EQUATION 3.62.

- Scenario B: for the second case a detector with a relative efficiency and perfect resolution is assumed (referred to as a detector with efficiency). The detector rejects some events from the data sample in accordance to the efficiency function shown in FIGURE 3.22. This efficiency is one up to $X_{\max} = 670$ g/cm^2 but above this value it falls linearly and becomes zero at $X_{\max} = 900$ g/cm^2. Notice that if in the parent distribution the proton fraction is 0.8, the proton fraction in the distribution recorded by the detector will not be 0.8 because the detector has a greater chance of detecting showers generated by iron which have smaller $X_{\max}$ than those generated by protons. In fact while the 99% of showers generated by iron nuclei are detected only the 65% of the showers generated by protons are actually recorded by the detector. Then, the detector introduces a bias in the proton fraction but this can be corrected for, using the correctly normalised likelihood (from EQUATION 3.65)

$$P(X_{\max}|\alpha) = \frac{\{\alpha g_p(X_{\max}) + (1-\alpha) g_{Fe}(X_{\max})\}\, \epsilon(X_{\max})}{\int_0^\infty \{\alpha g_p(X_{\max}) + (1-\alpha) g_{Fe}(X_{\max})\}\, \epsilon(X_{\max}) dX_{\max}}. \quad (3.73)$$

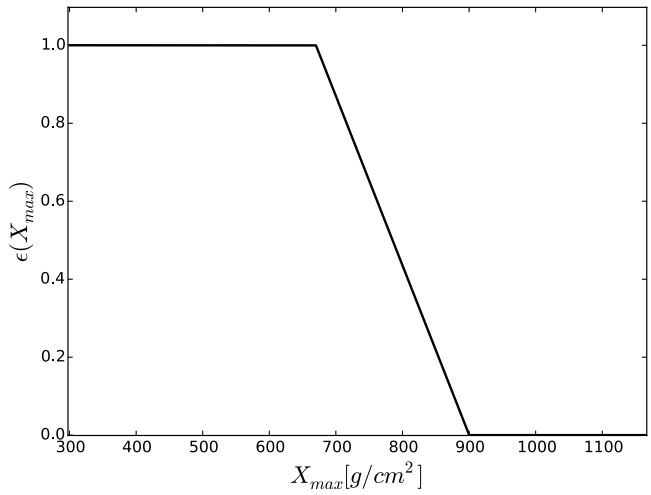


FIGURE 3.22: Relative efficiency of the detector in the example.

- Scenario C: the detector is assumed to have a limited resolution which has a "smearing" effect on the data. This is described as a probability density function of the difference between the actual $X_{\max}$ value and its measurement $X_{\max}^{obs}$. This function is shown in FIGURE 3.23 (a detector with efficiency and finite resolution, denoted by "real" or "realistic detector"). It is allowed to introduce a new bias in the parent data sample[9] by not being centred at zero.

[9]The parent data sample is which was detected by the ideal detector.

- Scenario D: finally the detector, besides having a finite resolution and limited efficiency, it is assumed to have a finite field of view covering only the $X_{\max}$ range between 650 g/cm^2 and 850 g/cm^2 (this is referred to as a "realistic detector with cut"). The resolution does not introduce any bias on the composition but the field of view introduces a new bias (added to the bias introduced by the efficiency) because only the 62% (56%) of proton(iron)-initiated showers are seen by the detector.

To infer the composition of the data samples obtained with these four detector models one must use Equation 3.68 for cases A,B and C and Equation 3.71 for case D.

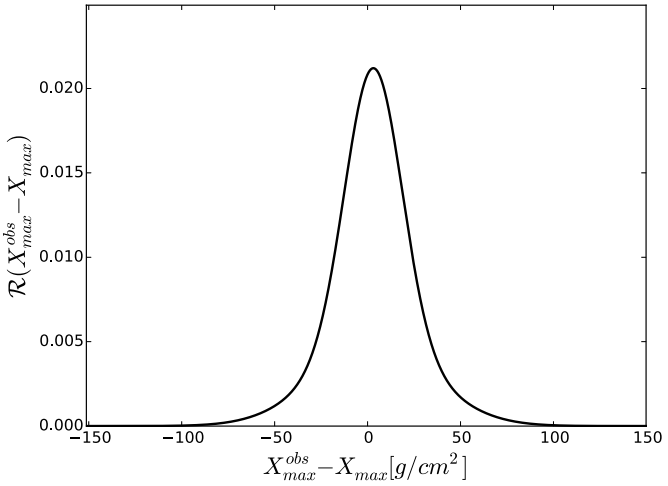
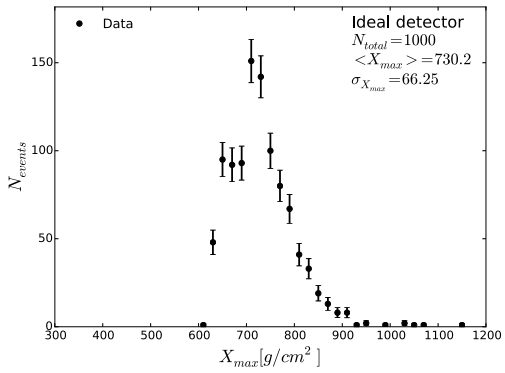
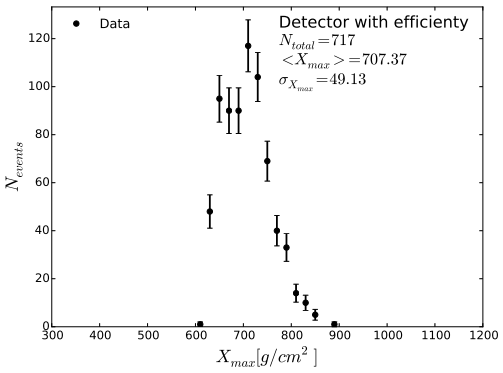


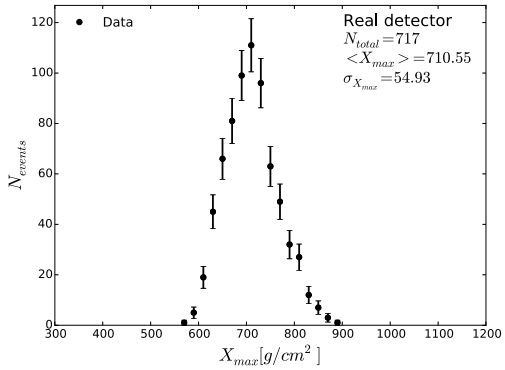
Figure 3.23: Response function of the detector in the example. It is not centred at zero.



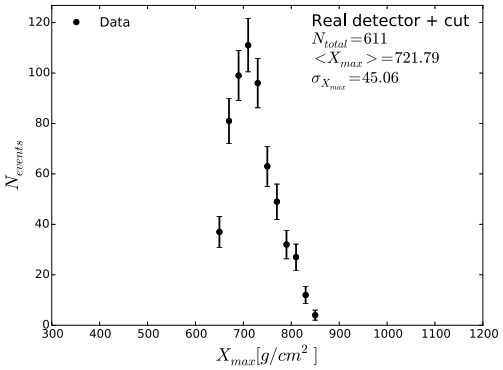
(A) Histogram of the $X_{\max}$ measured with an ideal detector.



(B) Histogram of the $X_{\max}$ data measured with a detector with an efficiency dependent of the actual value of $X_{\max}$.



(C) Histogram of the $X_{\max}$ data measured with a realistic detector.



(D) Histogram of the $X_{\max}$ data measured with a detector with finite field of view.

FIGURE 3.24: Data distributions for the different detectors:(A) perfect, (B) with efficiency, (C) realistic detector and (D) detector with a narrow field of view.

The obtained distributions are illustrated in FIGURE 3.24 for the four cases: A, B, C and D. Notice how the mean values and the standard deviations of the histograms change as different detector effects accumulate. The number of events of the data sample decreases as realistic features are included in the detector. This effects can be translated into a biased composition. However, using the correct normalised equations, the composition fraction can be inferred correctly. The posterior probabilities of the proton fraction for the different cases are shown in FIGURE 3.25 and the numerical values obtained for the proton fraction together the corresponding uncertainties as confidence intervals are listed in TABLE 3.4.

	α	C.I at 68%	C.I at 90%
Ideal detector	0.796	[0.781,0.810]	[0.772,0.820]
Det. with efficiency	0.795	[0.779,0.811]	[0.769,0.820]
Real detector	0.799	[0.781,0.817]	[0.769,0.828]
Real det. + FOV	0.808	[0.781,0.834]	[0.763,0.850]

TABLE 3.4: Estimated proton fraction and confidence intervals at 68% and 90% for the different detectors. The actual proton fraction is 0.8.

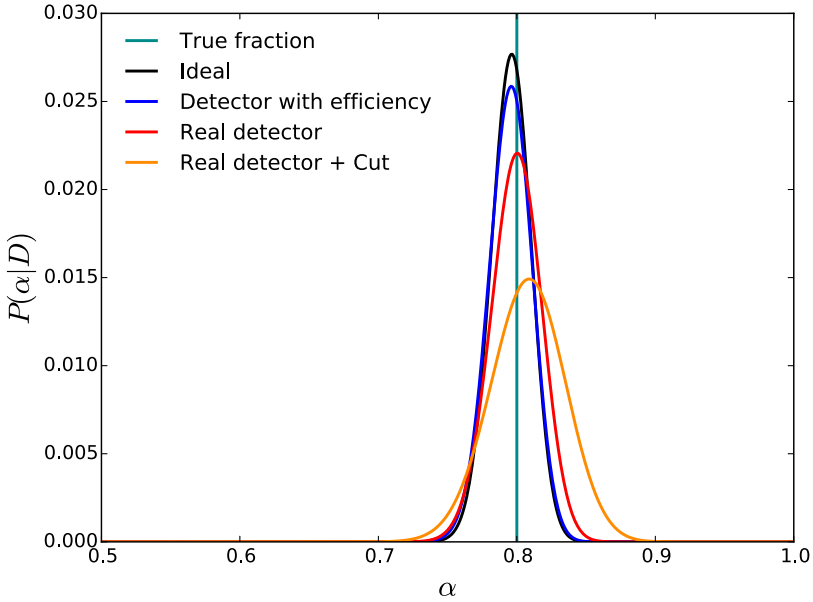


FIGURE 3.25: Posterior probability density functions for the different data samples seen by the different detectors: (black) ideal detector, (blue) detector with efficiency, (red) detector with efficiency and resolution and (orange) detector with efficiency, resolution and narrow field of view.

Once the composition fraction is inferred it is possible to build the posterior predictive distribution which provides a check on the inference. As a reminder, the posterior predictive distribution is the distribution of future events given the inferred composition fraction and it is given by

$$P(X_{\max}^{future}|D^{obs}) = \int_0^1 P(X_{\max}^{future}|\alpha)P(\alpha|D^{obs})d\alpha \tag{3.74}$$

This distribution can be compared with the data distribution. In FIGURE 3.26 the posterior predictive distributions are displayed given the inferred composition for

each detector model. The figure compares these predictive distributions directly to the observed data distributions. They are in indeed in very good agreement.

In spite of having used a detector with limited efficiency, finite resolution and biased because of a limited field of view (detector D) it is also possible to build the expected distribution of an ideal detector (detector A), or equivalently the actual distribution of events arriving to the Earth. This can be done combining the likelihood of the ideal detector (see EQUATION 3.62) with the $P(\alpha|D^{obs})$ given by the realistic detector with finite field of view (obtained using EQUATION 3.71) in EQUATION 3.74. As shown in FIGURE 3.27, the predicted distribution and the true distribution are also in very good agreement. This means that if the effects of the detector over the data are well known, then one can still estimate the true composition fraction even when the detector produces a selection bias which modifies the composition of the recorded data relative to the true composition.

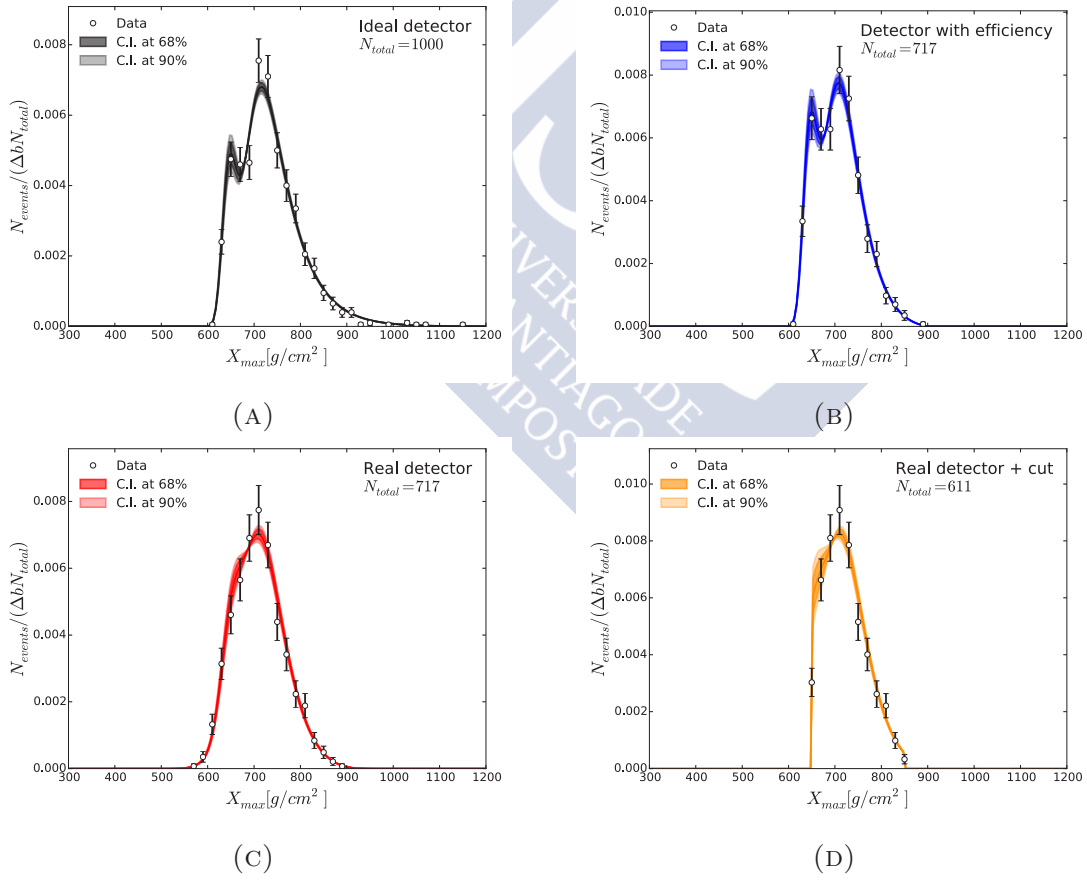


FIGURE 3.26: Posterior predictive distributions for the different detectors compared with the observed data by each detector:(A) perfect, (B) with efficiency, (C) realistic detector and (D) detector with a narrow field of view.

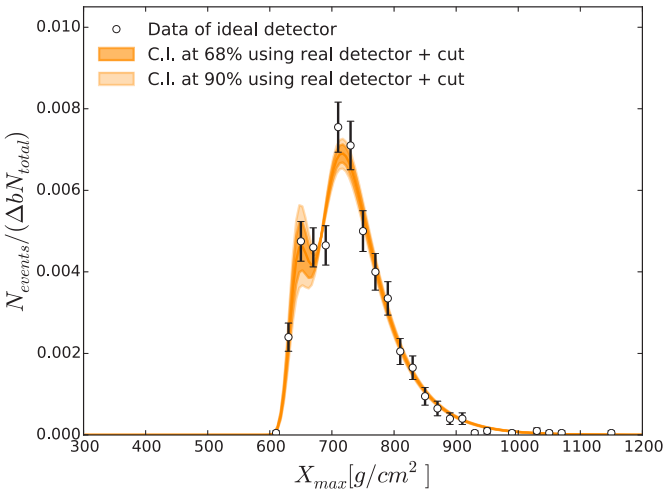


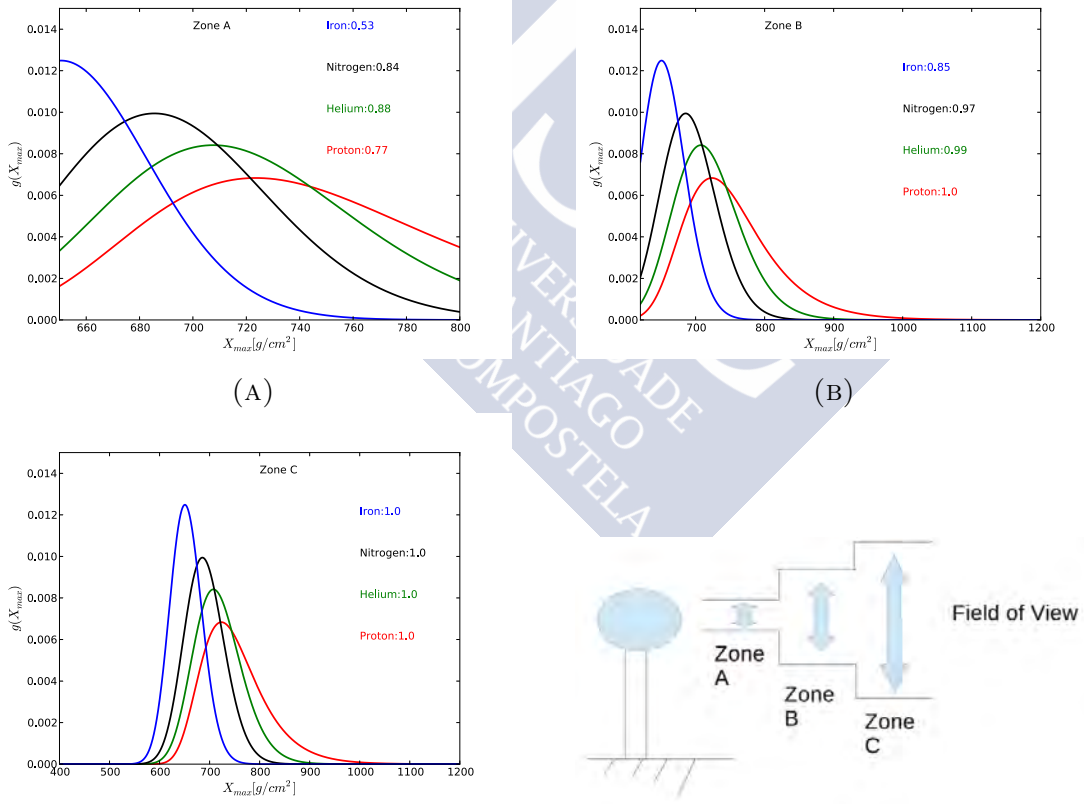
FIGURE 3.27: Posterior predictive distribution of data collected with a perfect detector given the fraction obtained with a realistic detector. It is compared with the data distribution recorded by the perfect detector.

3.7.2 The anti-bias cut versus using all data

Inspired by the approach used by the Pierre Auger Observatory we consider a simplified telescope shown at FIGURE 3.28 with three different zones. The nearest zone, "Zone A", the field of view of the telescope only covers a range in the shower development between 650 and 800 g/cm^2, the second zone, "Zone B", is beyond the "Zone A", and covers a range $620 - 1200$ g/cm^2. Finally, the far zone, "Zone C", covers all the depths. In FIGURE 3.28 the theoretical $X_{\max}$ distributions viewed by the detector are also shown for each zone. The resolution of the detector is taken to be the same as that of the previous example and the efficiency is considered to be 1 inside the field of view of the detector and 0 outside. Notice that analysing the events recorded in the "Zone A" with EQUATION 3.63 would lead to an estimated fraction which is biased. This analysis would not give the actual fraction of events arriving to the Earth but the fraction of recorded events by the telescope in this zone. The same applies to "Zone B". To correct this bias one can take two different approaches. The first one consists in applying an anti-bias cut. In this simplified example, the anti-bias cut consists on eliminating all the events detected in zones "A" and "B" and using only the events recorded in "Zone C" to perform the inference. In this example 2/3 of the events approximately would be lost (this is because of our simplified field of view). The other possibility would be to maintain all the detected events and use the correctly normalised EQUATION 3.71 in the Bayesian approach. In this case, the

region fulfilling the condition $\mathcal{C}$ is just the field of view of the telescope. The posterior probability function given an observed event $X^{obs}_{max,i}$ is

$$P(\alpha_p, \alpha_{He}, \alpha_N, \alpha_{Fe}|X^{obs}_{max,i}, I) = \\ = \left\{\alpha_P g_P(X^{obs}_{max,i}) + \alpha_{He} g_{He}(X^{obs}_{max,i}) + \alpha_N g_N(X^{obs}_{max,i}) + \alpha_{Fe} g_{Fe}(X^{obs}_{max,i})\right\} \times \\ \times P(\alpha_P, \alpha_{He}, \alpha_N, \alpha_{Fe}|I) \begin{cases} \frac{1}{0.77\alpha_P + 0.88\alpha_{He} + 0.84\alpha_N + 0.53\alpha_{Fe}} & \text{if event in “Zone A”} \\ \frac{1}{\alpha_P + 0.99\alpha_{He} + 0.97\alpha_N + 0.85\alpha_{Fe}} & \text{if event in “Zone B”} \\ 1 & \text{if event in “Zone C”} \end{cases} \tag{3.75}$$



(A) (B)

(C) (D)

FIGURE 3.28: In the panel (A) a simplified diagram of the telescope used in the simulations. In panels (B), (C) and (D) the fraction of the viewed $X_{\max}$ distributions for the different zones of the telescope are shown.

To give a quantitative example, 5000 different trials were performed with 2000 events per trial to study the inference of the composition fraction in the each of the three considered cases:

- using an ideal detector (infinite field of view, i.e, 2000 recorded events per trial)
- using events passing the anti-bias cut (only events in "Zone C") restricting the data set approximately to a third of the recorded events
- using all detected events but correcting for the biases with EQUATION 3.71

For each trial the primary fractions are varied randomly following a uniform distribution in composition space. The differences between the true generated fraction and the estimated fraction are shown in FIGURE 3.29 for each sample of trials. The estimated uncertainties at 68% of confidence level are shown in FIGURE 3.30. The numerical values for the average difference between the true generated composition and its inference in the three cases are shown in TABLE 3.5 together with the average of the uncertainty in each estimation. The results confirm that one can safely get rid of the bias introduced by the detector and recover the events which are cut when the anti-bias cut method is used. When the analysis of composition is performed using all the events with the correct normalised posterior the main achievement is that the uncertainties in the inferences are narrower than the inferences given using only the event passing the anti-bias cut, *i.e.*, the inferences are more accurate.

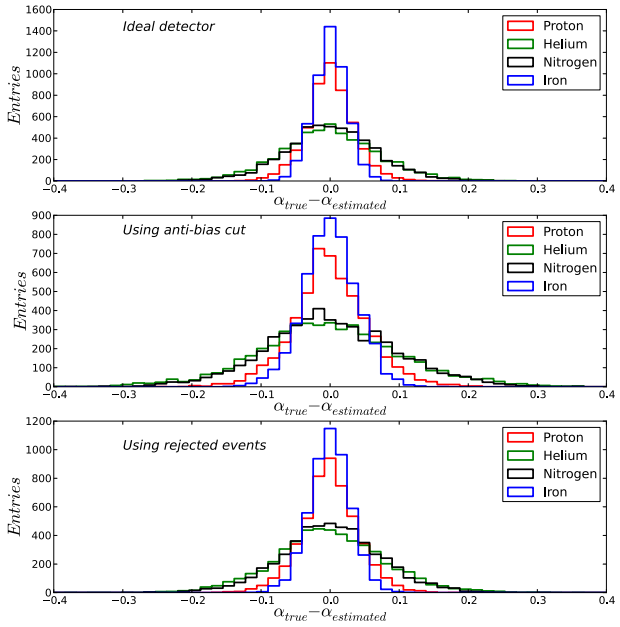


FIGURE 3.29: Differences between the true fractions and the estimated fractions. In the upper panel the differences are shown using all events and assuming a perfect detector. In the middle panel, only the events passing the anti bias cut are used for the analysis. In the lower panel all events are included for the analysis which accounts for the limited field of view of the detector.

	Ideal detector	Anti-bias cut	Using rejected events
$\langle \alpha_{true} - \alpha_{est} \rangle \, (Proton)$	0 ± 0.03	0 ± 0.05	0 ± 0.04
$\langle \alpha_{true} - \alpha_{est} \rangle \, (Helium)$	0 ± 0.07	0 ± 0.11	0 ± 0.08
$\langle \alpha_{true} - \alpha_{est} \rangle \, (Nitrogen)$	0 ± 0.07	0 ± 0.09	0 ± 0.07
$\langle \alpha_{true} - \alpha_{est} \rangle \, (Iron)$	0 ± 0.02	0 ± 0.04	0 ± 0.03
$\langle \sigma(\alpha) \rangle \, (Proton)$	0.03 ± 0.01	0.05 ± 0.02	0.04 ± 0.01
$\langle \sigma(\alpha) \rangle \, (Helium)$	0.07 ± 0.02	0.10 ± 0.03	0.08 ± 0.02
$\langle \sigma(\alpha) \rangle \, (Nitrogen)$	0.07 ± 0.02	0.09 ± 0.02	0.07 ± 0.02
$\langle \sigma(\alpha) \rangle \, (Iron)$	0.02 ± 0.01	0.04 ± 0.01	0.03 ± 0.01

TABLE 3.5: Mean values of the distributions $\alpha_{true} - \alpha_{estimated}$ and $\sigma(\alpha)$ for all the performed trials.

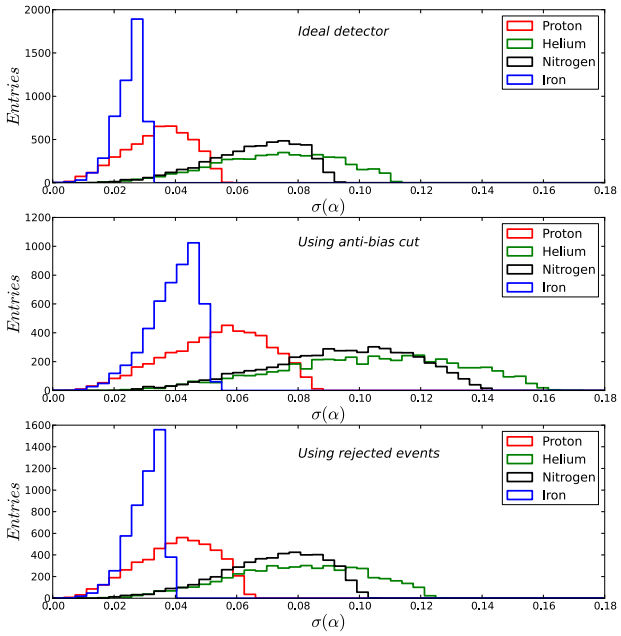


FIGURE 3.30: Standard deviations of the estimated fractions: using all events and assuming an ideal detector (upper panel), using only events passing the anti-bias cut (middle panel) and using all events and accounting for the limited field of view (lower panel).

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 4

The $X_{\max}$ data for the composition analysis

In this chapter the data for the composition analysis is presented. The data events are reconstructed with the official software of the Pierre Auger Observatory, the $\overline{\text{Off}}\underline{\text{line}}$ framework [63], and similarly the simulated events. The official description of the selection cuts for the composition analysis is presented. One of these cuts is the fiducial field of view cut (or anti-bias cut). This cut is done to ensure a bias-free composition of the $X_{\max}$ data distributions used for the analysis. Nevertheless, as was shown in Section 3.7, a unbiased distribution is not necessary to infer the true composition. Two data sets are used: one with the anti-bias cut applied and another one without the anti-bias cut. At the end of all selection cuts, we will use for the analysis data from December 2004 up to December 2012. The event with minimum energy has an energy of 0.6 EeV and the maximum energy is 79 EeV for events with anti-bias cut and 82 EeV for the data set without anti-bias cut.

We also introduce the numerical algorithm that will be used for the composition analysis of actual data in the next chapter. The treatment of the systematic uncertainties is also discussed.

4.1 Data selection

When an event has been recorded by the telescopes some conditions or cuts are mandatory to select the event for the composition analysis. These cuts can be divided

into two blocks: first a *pre-selection* which is followed by a *quality selection* to ensure a minimal distortion in the $X_{\max}$ distribution.

4.1.1 Pre-selection

The pre-selection is done to the set of candidate events to obtain a sample with a minimum quality requirements. Basically, these requirements consist in rejecting events that are not air showers produced by a cosmic ray (for example, lightning events) and checking that the surrounding conditions during the acquisition of the events are optimum and under control.

For the analysis of composition using the $X_{\max}$ data, only the events detected by the standard FD sites (Los Leones, Los Morados, Coihueco y Loma Amarilla) are accepted, rejecting the events detected by the HEAT telescope. There are also "laser events" produced artificially for monitoring the aerosol content of the atmosphere with the laser facilities XLF and CLF. These events must be rejected. Events detected within periods with reported problems of the FD operation are also rejected. If the pixels of the cameras were calibrated with bad parameters the event is rejected as well as if the PMT camera has saturated pixels because this could induce errors in the reconstruction of the shower profile. When the reconstruction of the longitudinal shower profile fails the e events are also rejected. When the shutters of the telescopes are closing the background light becomes smaller but events can be still detected. A minimum standard deviation in the photon counting is required to accept events recorded during the closing of the shutters.

Another kind of preselection cuts refer to the tanks. To ensure that the event corresponds with an extensive air shower and not with other kind of event (for example, lightning) at least one triggered tank is mandatory. Moreover, a tank could be triggered by coincident atmospheric muons and these events are rejected. The reconstructed shower core must be near the triggered station (less than 1.5 km) in order to guarantee a good geometry reconstruction. The time of the event is recorded using a GPS with a timing of one pulse-per-second. Nevertheless, a fast oscillator can measure fractions of microseconds with a frequency of 10 MHz. This oscillator has inaccuracies and this inaccuracies must be corrected for. Events without this correction are rejected.

The latest group of preselection cuts is related to the atmospheric conditions. First of all, the showers are accepted only if the atmospheric conditions were being monitored during their detection. Only events with a integrated vertical aerosol depth (VAOD) from the ground to 3 km which are below than 0.1 are accepted. That way we ensure a poor air contamination due to aerosols. Finally, using the laser facilities (XLF and CLF), the LIDARs and the cloud monitoring cameras, we can avoid possible distortions in the reconstruction of the shower due to the presence of clouds (see FIGURE 4.1 for a schematic picture of the system).

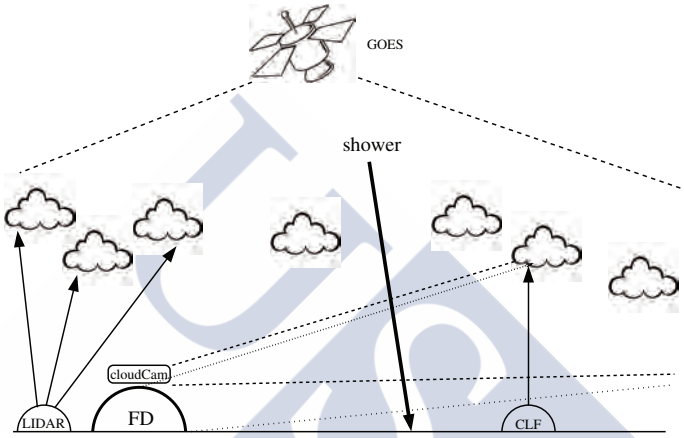


FIGURE 4.1: Illustration of the cloud monitoring. Figure taken from [64].

The event is accepted if no clouds are observed with any of the cloud monitoring devices. In the presence of clouds, the event is accepted if the position of the clouds measured by LIDARs and lasers is either outside of the geometrical field of view of the telescope or it is 400 g/cm^2 above the fiducial depth cut (explained below in the text). Finally, if the information to check the explained conditions is not available the event is rejected when the average cloud fraction measured by LIDARs is greater than 25%.

4.1.2 Quality selection

Once the pre-selection has been done the next step is to cut the data so that a minimum quality can be guaranteed for composition studies. This process has to be carefully done to avoid introducing biases in the $X_{\max}$ measurements.

A minimum energy is required ($\log_{10}(E/\text{eV}) \geq 17.8$) to accept the event for the analysis. During the pre-selection, only events with at least one triggered station and

a maximum tank-core distance of 1.5 km have been selected. The trigger probability is dependent on composition and therefore the cut implies a selection bias which favours one composition over another. By choosing a high value for the triggering probability we ensure that the residual bias is low. Then a trigger probability of the tanks within 1.5 km of the shower core above 95% is mandatory, irrespectively of the primary composition being proton or iron.

The value of the $X_{\max}$ is obtained fitting a Gaisser-Hillas profile to the observed fluorescence light. Some more cuts are performed to ensure that the quality of the $X_{\max}$ measurement is good enough.

- For a given event, it is possible that the reconstructed $X_{\max}$ falls outside the range of slant depths that is accessible by the telescopes. The current FD telescopes cover elevation angles between 1.5° and 30°. For a given cosmic ray "geometry" (here we refer to geometry as the arrival direction and the impact point which determine the axis of the shower) these angles intersect the shower trajectory at points which mark the minimum and maximum slant depths that can be observed with a given telescope (called field of view of the telescope for the event). Events that have $X_{\max}$ outside the field of view are thus removed because it cannot be guaranteed that the $X_{\max}$ reconstruction is accurate enough.

- The reconstruction of the maximum of the shower profile is a complicated function of many variables such as the geometry of the shower, the energy and the atmospheric conditions. The "expected precision" is defined replacing the observed profile of the shower by a Gaisser-Hillas profile for the same energy and maximum depth and propagating the uncertainties of the expected triggered number of pixels to an uncertainty of $X_{\max}$. Events with this expected precision equal or larger than 40° are discarded.

- Showers which are observed with an observation angle below 20° are also removed. The direction of these showers is quite close to the observation direction. These showers can have a large Cherenkov contamination and the $X_{\max}$ reconstruction can lead to large systematic errors. In addition the geometrical reconstruction of these events can also have large errors.

The expected precision and the observation angle define a range $[X_{\text{low}}, X_{\text{up}}]$ for each shower called the expected field of view. The shower is accepted if the expected field of view is large enough to accommodate the bulk of the $X_{\max}$ distribution.

Fiducial cuts

The measurement of the $X_{\max}$ using just the selection criteria mentioned above is subject to biases. Showers with a given geometry will be accepted or rejected for the analysis depending on their $X_{\max}$ position in relation to the expected field of view. Since different primary particles have different $X_{\max}$ distributions it is clear that the accepting criteria for showers will necessarily be composition dependent. As a result the distribution of observed values of $X_{\max}$ will have a composition mixture that does not necessarily reflect the composition of cosmic rays when they arrive at the Earth. This is one of the most important experimental problems in addressing composition measurements with the FD.

The Pierre Auger collaboration has devised a set of cuts, often referred to as *fiducial cuts*, specifically designed to minimise the composition bias in the measurement of the average $\langle X_{\max} \rangle$ (see [65]). The idea is to select events using their energy, arrival direction and impact point alone and to determine if the geometry is "acceptable". An event is accepted if we can ensure that by measuring events of the same energy and geometry we could reproduce the average value of $X_{\max}$ irrespectively of the unknown composition of the cosmic rays.

The fiducial cuts (or *anti-bias cuts*), are performed by defining a new range for each shower $[X^{\text{fid}}_{\text{low}}, X^{\text{fid}}_{\text{up}}]$, called *fiducial field of view*, established empirically by studying how does the mean value of the measured $X_{\max}$ change when the values of the parameters $[X^{\text{fid}}_{\text{low}}, X^{\text{fid}}_{\text{up}}]$ are varied in turn within fixed energy bins. The actual values are chosen demanding that the mean is no altered by more than 5% because of each of these cuts. The values chosen for the analysis given by the collaboration ([66]) are given by the following characterisation:

$$X^{\text{fid}}_{\text{low,up}}(E) = \begin{cases} p_1 & \text{if } \log_{10}(E/\text{eV}) > p_3 \\ p_1 + p_2(\log_{10}(E/\text{eV}) - p_3)^2 & \text{if } \log_{10}(E/\text{eV}) \leq p_3 \end{cases} \tag{4.1}$$

Here, $\vec{p}_{\text{low}} = (695.7, -34.6, 19.8)$ and $\vec{p}_{\text{up}} = (891.8, -186.3, 18.2)$. The effects of these cuts have also been checked with showers simulations allowing any composition ranging between proton and iron a several hadronic models. This simulations indicate

that the chosen cuts do no significantly alter the mean and variance of the measured events. Then, selecting events with a fiducial field of view contained within the expected field of view it is guaranteed that $\langle X_{\max} \rangle$ is not significantly altered and it is inside the "safe" region with the acceptable expected precision.

Finally, three more cuts are applied which are directly related with the quality of the Gaisser-Hillas fit. First, events with gaps in their track length larger than 20% of the total track length of the shower are rejected. In order to avoid outliers to keep only high quality showers a "standard-normal" transformation to the χ^2 fit is performed

$$z = \frac{\chi^2 - \text{ndof}}{\sqrt{2\text{ndof}}} \tag{4.2}$$

rejecting events whose z deviates more than 2.2 standard deviations relative to the mean of the z distribution. The last cut consists in accepting showers with a minimum observed track length of 300g/cm^2.

For composition analysis we are going to work with two data samples to compare results. In the default analysis we apply all the cuts explained above getting 19749 events. For the second sample the anti-bias cut is removed obtaining a total of 44218 events for the same data period. Note that removing the anti-bias cut we gain over a factor 2 in the number of events.

4.2 Detector description

For the analyses that will follow we need to have a good description of the detector, its efficiency and response functions. This description has to be done in two scenarios, when all the cuts accounted for and when the fiducial cuts are removed.

The development of models to provide this description is a complex task and falls beyond the scope of this thesis. Instead we will use the available documentation about it that has been obtained within the collaboration. A summary of the efficiency and the response functions using the fiducial cuts can be found in [67]. It is also necessary to account for the systematic uncertainties. A detailed description of these can be found in [64]. Here we are going to prepare and collect all the functions that are needed for these descriptions.

4.2.1 Detector using the fiducial cut

Efficiency

The relative acceptance of the detector has been studied as a function of the true $X_{\max}$ and energy using simulations. It is defined as the ratio of selected to generated events. FIGURE 4.2 displays the relative efficiency with the systematic uncertainties for two different energies. It can be parameterized as:

$$\epsilon_{rel}(X_{\max}, E) = \begin{cases} e^{\frac{X_{\max}-x_1(E)}{\lambda_1(E)}} & \text{if } X_{\max} \leq x_1 \\ 1 & \text{if } x_1 < X_{\max} \leq x_2 \\ e^{-\frac{X_{\max}-x_2(E)}{\lambda_2(E)}} & \text{if } X_{\max} > x_2 \end{cases}. \tag{4.3}$$

The relative efficiency depends on the energy through the parameters $x_{1,2}$ and $\lambda_{1,2}$. These parameters can be given as continuous functions of the shower energy as:

$$x_1(E) = 595\{^{585}_{604.9}\} + 50.1\{^{20.9}_{78.9}\} \log_{10}(E/\text{EeV}) - 60.7\{^{62.2}_{58.3}\} \log^2_{10}(E/\text{EeV}), \tag{4.4}$$

$$\lambda_1(E) = 146\{^{163.7}_{127.7}\} + 247\{^{260.6}_{232.5}\} \log_{10}(E/\text{EeV}) - 16.8\{^{22.8}_{56.5}\} \log^2_{10}(E/\text{EeV}), \tag{4.5}$$

$$x_2(E) = 884\{^{892}_{876}\} + 18.2\{^{19.1}_{17.1}\} \log_{10}(E/\text{EeV}), \tag{4.6}$$

$$\lambda_2(E) = 104\{^{110.9}_{96.9}\} + 61.1\{^{62}_{60}\} \log_{10}(E/\text{EeV}). \tag{4.7}$$

Here, the bracketed numbers are the upper and lower limits of the parameters and they are used to obtain a systematic uncertainty in the composition analysis that will be made. The above functions and all parameters are taken from [64].

Response

The final response function can be described as function of the difference $X^{rec}_{\max} - X^{true}_{\max}$ which takes the form of a mixture of two Gaussian distributions parameterized in terms of f, σ_1 and σ_2 all of which are energy dependent [64].

$$\mathcal{R}(X^{rec}_{\max} - X^{gen}_{\max}, E) = fG(\mu, \sigma_1) + (1-f)G(\mu, \sigma_2) \tag{4.8}$$

The functions $f(E)$, $\sigma_1(E)$ and $\sigma_2(E)$ are tabulated in terms of the energy in TABLE 4.1. These values are extracted from [67]. These values take into account all effects that contribute to the response function (detector effects, atmosphere and

energy scale uncertainty) and have been used for the official composition analysis [66].

In reference [64] alternative functions for $f(E)$, $\sigma_1(E)$ and $\sigma_2(E)$ are constructed and parameterised but taken only into account the detector effects. The characterisation in [64] is given in terms of $f(E)$ and two equivalent functions: $\sigma^2_{full}(E)$ and $\alpha(E) = \sigma_2/\sigma_1$. It is straightforward to obtain σ_1 and σ_2 by solving the system

$$\begin{cases} \sigma^2_{full} & = f\sigma_1^2 + (1-f)\sigma_2^2 \\ \alpha & = \frac{\sigma_2}{\sigma_1} \end{cases} \tag{4.9}$$

We will obtain $\sigma^2_{full}(E)$ and $\alpha(E)$ in a similar way to [64] but using the results of Table 4.1.

$\log(E)$ range	σ_1	σ_2	f
$[17.8, 17.9)$	17.5 ± 0.7	33.7 ± 1.4	0.62
$[17.9, 18.0)$	16.7 ± 0.7	32.9 ± 1.4	0.63
$[18.0, 18.1)$	15.9 ± 0.7	31.9 ± 1.4	0.63
$[18.1, 18.2)$	15.1 ± 0.7	31.0 ± 1.4	0.64
$[18.2, 18.3)$	14.4 ± 0.7	30.0 ± 1.4	0.65
$[18.3, 18.4)$	13.8 ± 0.7	29.1 ± 1.5	0.66
$[18.4, 18.5)$	13.3 ± 0.7	28.1 ± 1.6	0.67
$[18.5, 18.6)$	12.8 ± 0.8	27.1 ± 1.6	0.68
$[18.6, 18.7)$	12.3 ± 0.8	26.3 ± 1.7	0.69
$[18.7, 18.8)$	12.0 ± 0.8	25.4 ± 1.8	0.70
$[18.8, 18.9)$	11.7 ± 0.9	24.7 ± 1.9	0.70
$[18.9, 19.0)$	11.5 ± 0.9	24.1 ± 1.9	0.71
$[19.0, 19.1)$	11.3 ± 0.9	23.6 ± 1.9	0.72
$[19.1, 19.2)$	11.2 ± 0.9	23.3 ± 2.0	0.73
$[19.2, 19.3)$	11.1 ± 0.9	23.1 ± 2.0	0.74
$[19.3, 19.4)$	11.1 ± 1.0	23.1 ± 2.0	0.75
$[19.4, 19.5)$	11.1 ± 1.0	23.2 ± 2.0	0.76
$[19.5, \infty)$	11.2 ± 1.0	23.7 ± 2.1	0.77

Table 4.1: Parameters of the $X_{\max}$ resolution with their systematic uncertainties extracted from [67]

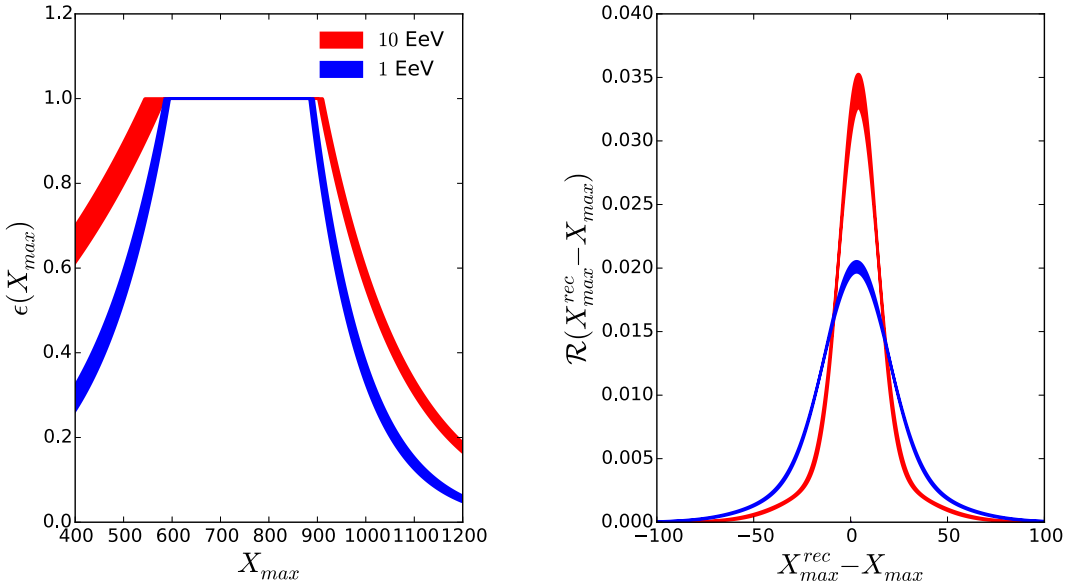


FIGURE 4.2: Relative acceptance (left) and response function (right) for two different energies: 1 EeV (blue) and 10 EeV (red) for events passing the fiducial cut. The width of the bands represents the systematic uncertainties.

The relative weight of the two Gaussians that is obtained is given by:

$$f(E) = 0.63 + 0.088 \log_{10}(E/\text{EeV}). \tag{4.10}$$

The ratio of the Gaussians widths , $\alpha = \sigma_2/\sigma_1$, can be expressed as:

$$\alpha(E) = 2.1 + 0.71 \log_{10}(E/\text{EeV}). \tag{4.11}$$

Finally, the total width σ_{full} of the response function and the bias μ (mean of the response function) are given by:

$$\mu(E) = -3.4 + 0.93 \log_{10}(E/\text{EeV}) + b_{LW_{corr}}, \tag{4.12}$$

$$\sigma^2_{full}(E) = f\sigma^2_1 + (1-f)\sigma^2_2 = 14.9^2\{^{16.2^2}_{13.7^2}\} + \left[18.2\{^{18.5}_{17.9}\}e^{y(E)}\right]^2 \tag{4.13}$$

where $b_{LW_{corr}}$ and $y(E)$ are given by EQUATION 4.14 and EQUATION 4.15 respectively.

$$b_{LW_{corr}}(E) = \frac{6.5}{1 + \exp\left(\frac{\log_{10}(E) - 18.23}{0.41}\right)} \tag{4.14}$$

$$y(E) = -\frac{\log_{10}(E/\text{EeV})}{1.48\{^{1.44}_{1.51}\} - 0.65\{^{0.59}_{0.71}\}\log_{10}(E/\text{EeV})} \tag{4.15}$$

The parameter $b_{LW_{corr}}$ takes into account the small bias introduced by the algorithm to measure the lateral width correction when it is applied to simulated showers (see [64] and [68]). An example of the response function for two different energies is also shown in the right panel of FIGURE 4.2. The total uncertainties in the response function are dominated by the energy scale uncertainty [69] which is of order 14%.

4.2.2 Detector without fiducial cut

There are no publications about the behaviour of the detector without the fiducial cuts. Nevertheless some work has been done in relation with this issue which is reflected in internal notes of the collaboration , see for example [70]. Most of the information that we will use here has been obtained through [71].

Efficiency

As in the previous section, the relative efficiency is given by simulations taking the ratio of selected showers to generated showers. It is parameterised with the same functional form as that used for the events passing the anti-bias cut (EQUATION 4.3). The new values of the parameters $x_{1,2}$ and $\lambda_{1.2}$ are given by EQUATIONS 4.16-4.19.

$$x_1(E) = 564.6\{^{574.6}_{547.8}\} - 177.9\{^{163.3}_{166}\}\ \log_{10}(E/\mathrm{EeV}) + 8.1\{^{-14.3}_{-7.2}\}\ \log^2_{10}(E/\mathrm{EeV}), \quad (4.16)$$

$$\lambda_1(E) = 157.3\{^{141.2}_{170.3}\} + 357\{^{348.6}_{415.8}\}\ \log_{10}(E/\mathrm{EeV}) - 44.4\{^{-58.2}_{111.9}\}\ \log^2_{10}(E/\mathrm{EeV}), \quad (4.17)$$

$$x_2(E) = 816.7\{^{820.4}_{818.4}\} - 96.1\{^{99.5}_{98.5}\}\ \log_{10}(E/\mathrm{EeV}) + 24.6\{^{25.9}_{25}\}\ \log^2_{10}(E/\mathrm{EeV}), \quad (4.18)$$

$$\lambda_2(E) = 227.6\{^{228.1}_{224.1}\} + 142.1\{^{135.1}_{159.3}\}\ \log_{10}(E/\mathrm{EeV}) - 58.5\{^{52.6}_{69.4}\}\ \log^2_{10}(E/\mathrm{EeV}). \quad (4.19)$$

An example for two different energies is shown on the left panel of FIGURE 4.3. Comparing FIGURES 4.2-4.3 one can see that the efficiency is different if the fiducial cut is selected or not, as could be anticipated.

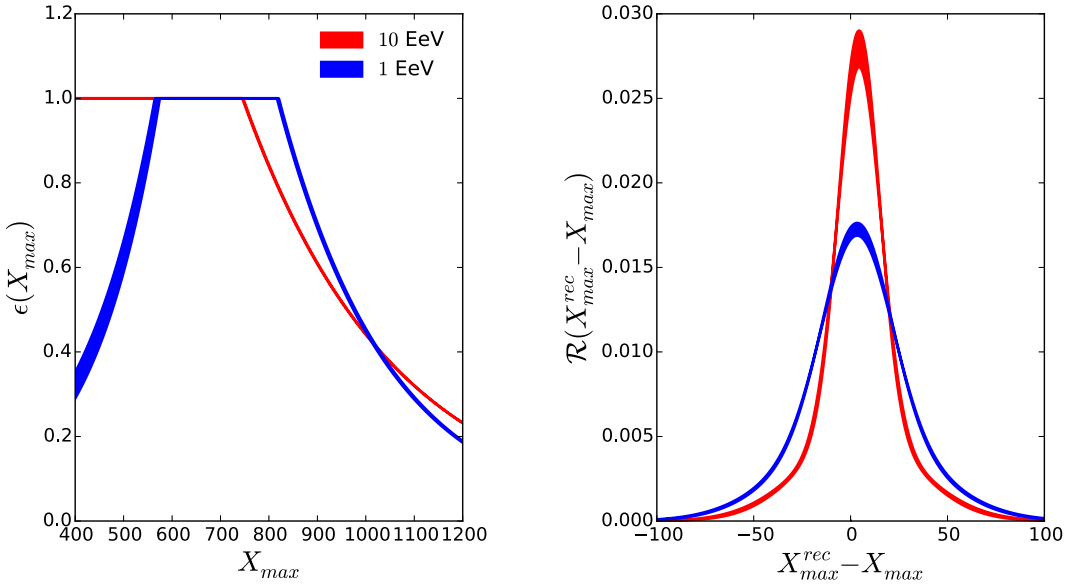


FIGURE 4.3: Relative acceptance (left) and response function (right) for two different energies: 1 EeV (blue) and 10 EeV (red)for a data set without fiducial cuts. The width of the bands denotes the systematic uncertainty.

Response

The information about the response function for this kind of events is very limited. The width of the response function is the combined result of different effects grouped into three categories: those due to the detector itself, those due to the aerosol content in the atmosphere and those due to molecular atmospheric dispersion. The statistical uncertainty of the aerosol content contributes to the resolution of the $X_{\max}$ and is referred to as the contribution from aerosols. The uncertainty due to molecular dispersion can be attributed to the precision to which the density profiles of the atmosphere are known as a function of height.

The "detector contribution" describes all effects in the reconstruction which are not due to any of the previous contributions (atmospheric and aerosols). Although it has not been published it has been parameterised with and without the fiducial cuts [71].

To calculate the full detector response without fiducial cuts we first obtain atmospheric effects subtraction the detector response form the full response for the data set with fiducial cuts:

$$\sigma^2_{atm} = \sigma^2_{full} - \sigma^2_{det}, \tag{4.20}$$

where σ^2_{atm} is the sum of the aerosol and molecular contributions. Although this result is obtained with fiducial cuts we can assume that the atmospheric contribution is the

same for the data without fiducial cuts. With this assumption it is straightforward to calculate the full width using again EQUATION 4.20

The new response function for events without fiducial cuts is parameterised using EQUATION 4.8 again. The functions are now expressed as:

$$f(E) = 0.51 + 0.11 \log_{10}(E/\mathrm{EeV}), \tag{4.21}$$

$$\alpha(E) = \frac{\sigma_2(E)}{\sigma_1(E)} = 2.1 + 0.73 \log_{10}(E/\mathrm{EeV}), \tag{4.22}$$

$$\mu(E) = -3.1 + 0.98 \log_{10}(E/\mathrm{EeV}) + \frac{6.5}{1 + \exp\left(\frac{\log_{10}(E) - 18.23}{0.41}\right)}, \tag{4.23}$$

which already include atmospheric contributions as explained above. The values of σ_1 and σ_2 are again obtained using EQUATION 4.9.

The remaining open question is: how can we get the systematic uncertainty of σ_{full} for the events without fiducial cuts. We will again rely on results obtained for data with fiducial cuts. The relative uncertainties of σ_1 and σ_2 obtained from TABLE 4.1 are plotted in FIGURE 4.4 which clearly indicates that the relative uncertainties of σ_1 and σ_2 are equal. We will assume that this relation hold for events without fiducial cuts. Once the absolute value of σ_1 and σ_2 are found, we can obtain the systematic uncertainty and propagate it to σ_{full}

We obtain the following characterisation of the width of the response function for the events without anti-bias cut:

$$\sigma^2_{full}(E) = 16.4^2\{^{17.3^2}_{15.2^2}\} + \left[21.7\{^{22.5}_{21.1}\}e^{y(E)}\right], \tag{4.24}$$

where $y(E)$ is given by

$$y(E) = -\frac{\log_{10}(E/\mathrm{EeV})}{2.13\{^{2.24}_{2.06}\} - 0.69\{^{0.58}_{0.77}\}\log_{10}(E/\mathrm{EeV})}. \tag{4.25}$$

FIGURE 4.5 gives a complete summary of the response function with and without fiducial cuts. The different resolutions due to the detector, molecular + aerosols and the full resolution are shown together for both data samples with the corresponding systematic uncertainties.

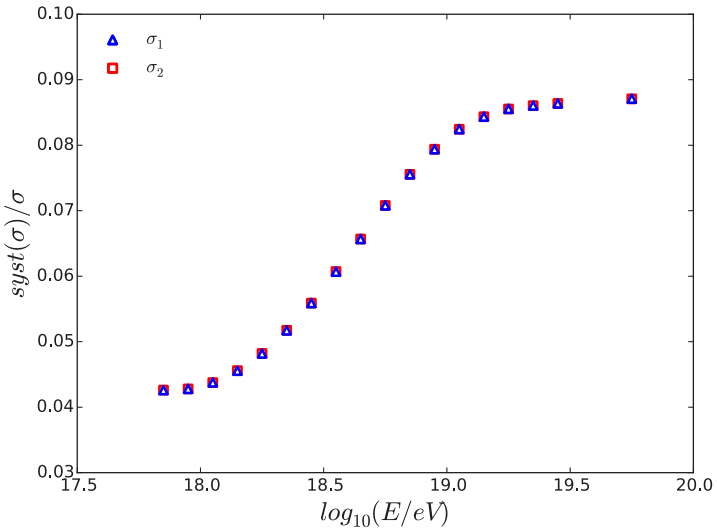


FIGURE 4.4: Ratio between the systematic uncertainties in $\sigma_{1,2}$ and the absolute value of $\sigma_{1,2}$ for a data sampling passing the fiducial cuts. Data extracted from [67].

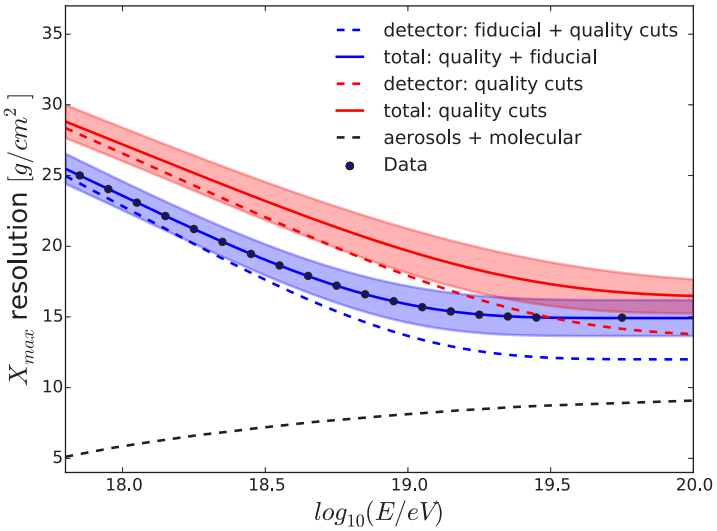


FIGURE 4.5: The black points are the value of σ_{full} for the events with anti-bias cut extracted from TABLE 4.1. The blue continuous line is the fit to these points and the blue band represents the systematic uncertainty. The blue(red) dashed line is the detector contribution to the resolution of $X_{\max}$ for the events with(without) anti-bias cut. The black dashed line is the contribution to the $X_{\max}$ resolution which is not from the detector. The red continuous line is the total width of the response function for the events without fiducial cut and the red band is its uncertainty obtained as explained in text.

4.3 Data description

The number of events as a function of the energy is shown in FIGURE 4.6 in energy bins $\Delta\log_{10}(E/\text{eV}) = 0.1$. The number of events in the data sample without fiducial cuts more than doubles the number of events in the data with them and for this reason we expect to get a more accurate inference for the composition using the events without fiducial cuts. Unfortunately the increase in the number of events is larger at low energies and we do not expect a significant improvement in the composition inference at energies above $10^{19.5}$ eV where the flux suppression occurs.

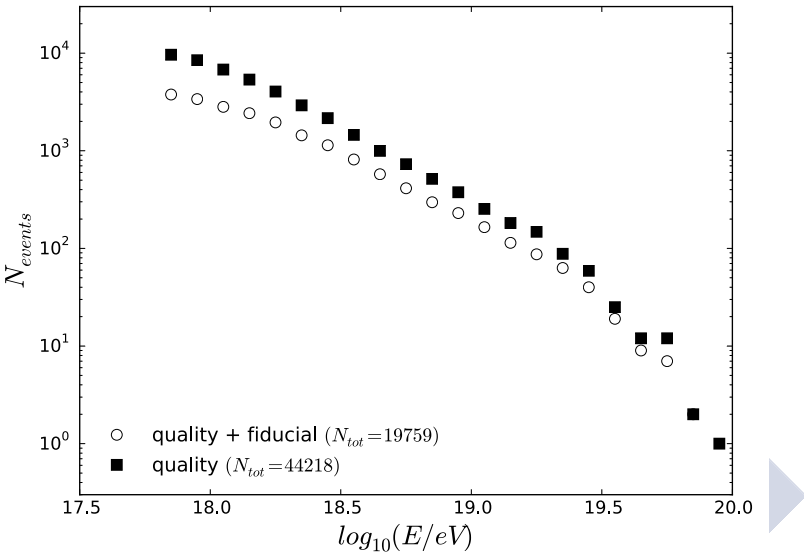


FIGURE 4.6: Number of events as a function of the energy for the composition analysis. The events passing the fiducial cut are denoted by opened circles and the events without fiducial cut are denoted by black squares.

We recall that the distributions without field of view cuts are expected to have a bias since no cut has been made to minimise the effect. As a result we expect different $X_{\max}$ distributions for the different data samples and also different moments. The mean value and the standard deviation of the observed $X_{\max}$ distribution for both data samples are shown in FIGURE 4.7 as a function of energy. Indeed we note an effect in the newly obtained means without fiducial cuts, which are lower than those that have been published except for the first energy bin ($17.8 \leq \log_{10}(E/\text{eV}) < 17.9$. No significant difference can be appreciated in the standard deviations except for the lowest energy bin again. The corresponding $X_{\max}$ histograms are shown in FIGURES 4.8-4.10.

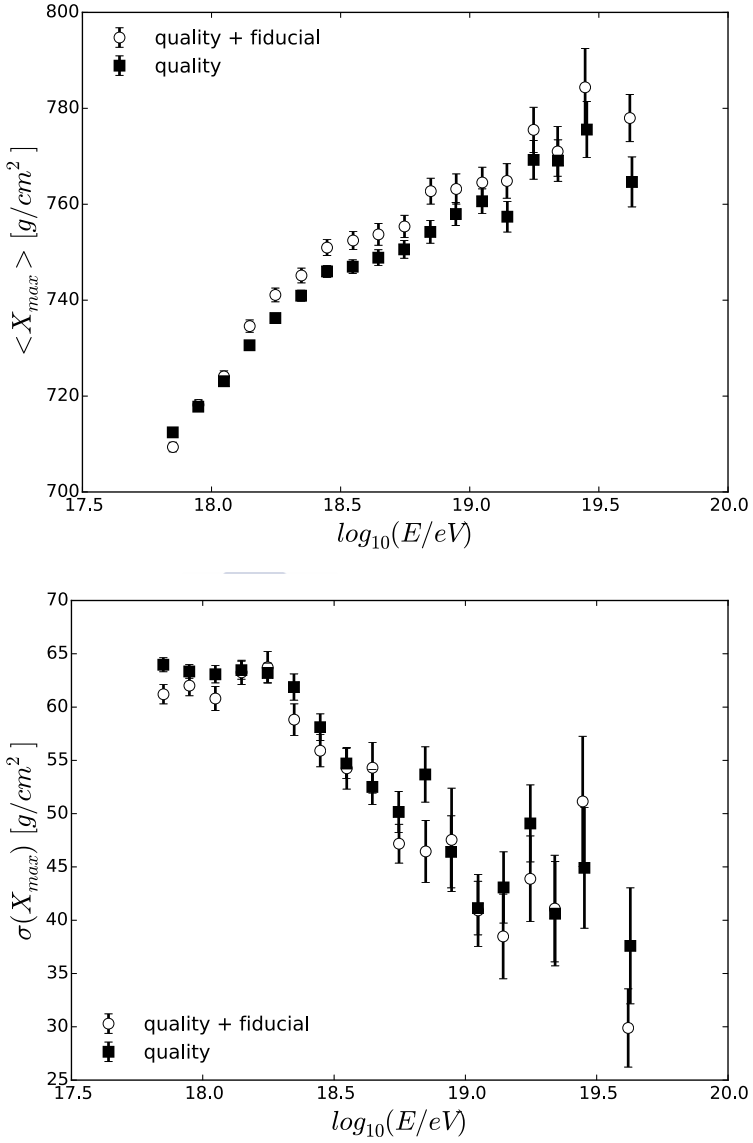


FIGURE 4.7: Comparison of $\langle X_{\max} \rangle$ (upper panel) and $\sigma(X_{\max})$ (lower panel) for the events with (open circles) and without (black squares) anti-bias cut. The energy bins correspond to $\Delta \log_{10}(E/\mathrm{eV}) = 0.1$ with the exception of the last bin which has been taken to include all events with $\log_{10}(E/\mathrm{eV}) \geq 19.5$.

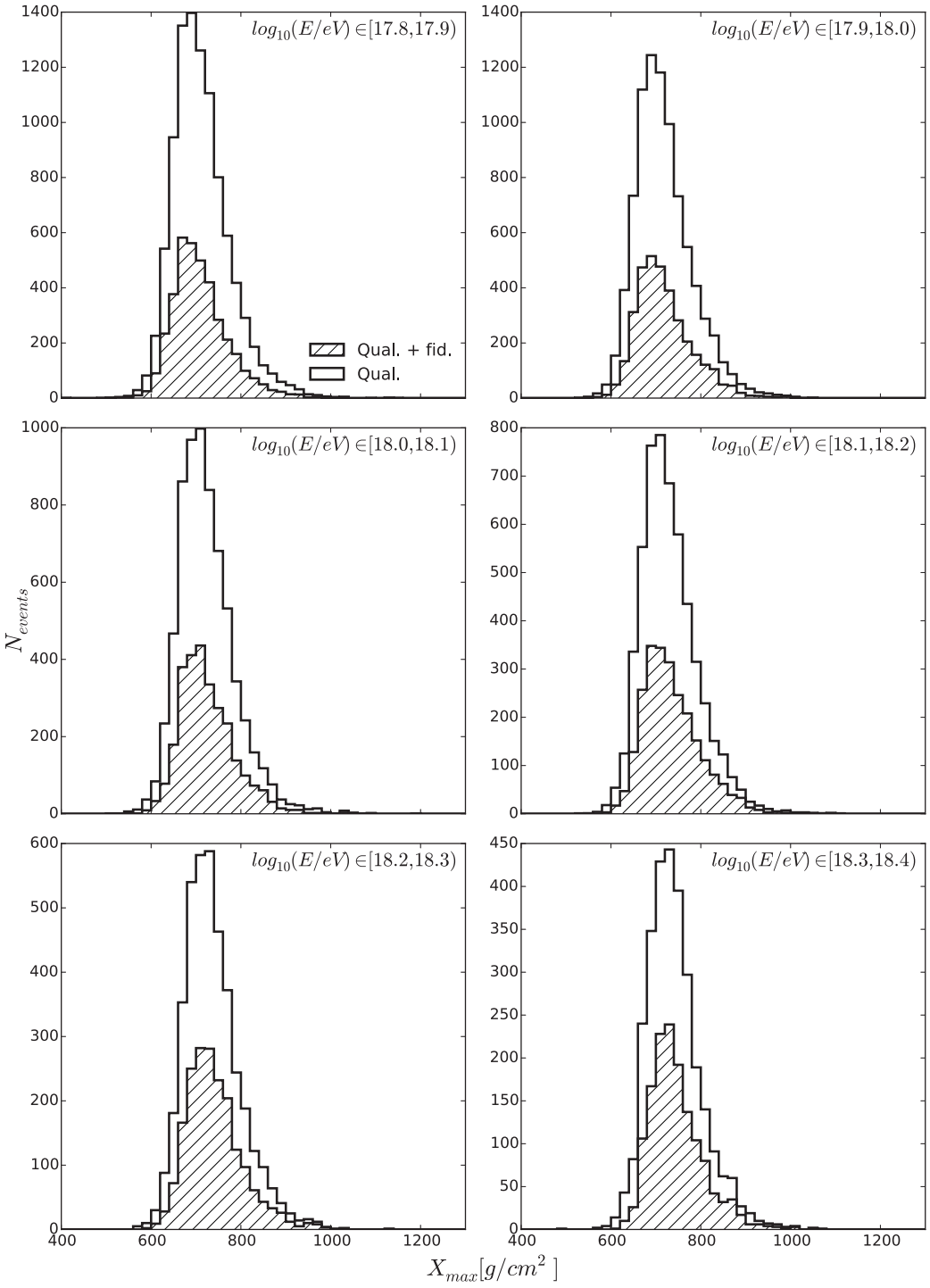


FIGURE 4.8: Comparison of $X_{\max}$ data histograms used for the composition analysis with (hatched) and whiteout (empty) as labelled.

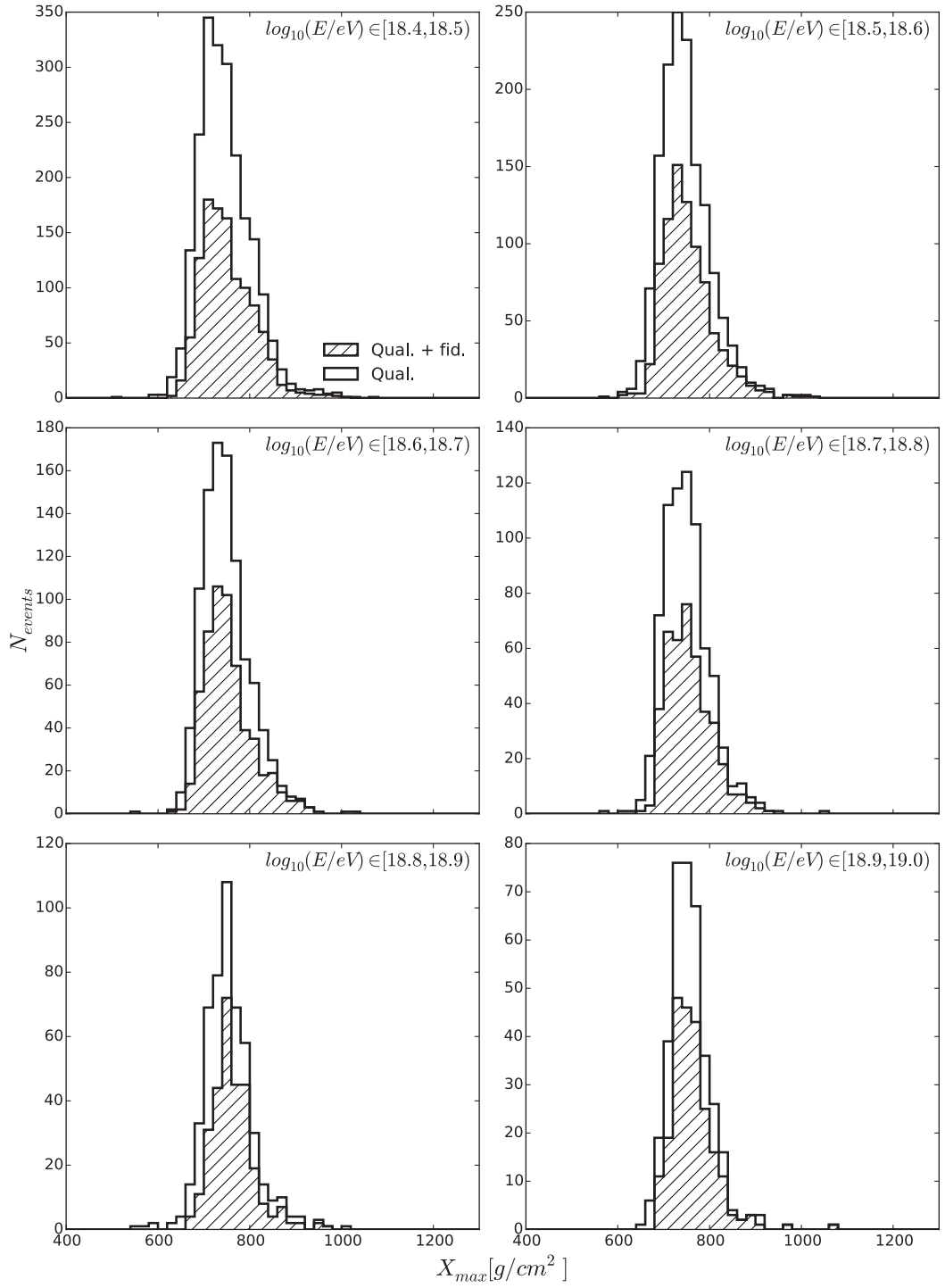


FIGURE 4.9: Same as FIGURE 4.8 but different energy ranges as labelled.

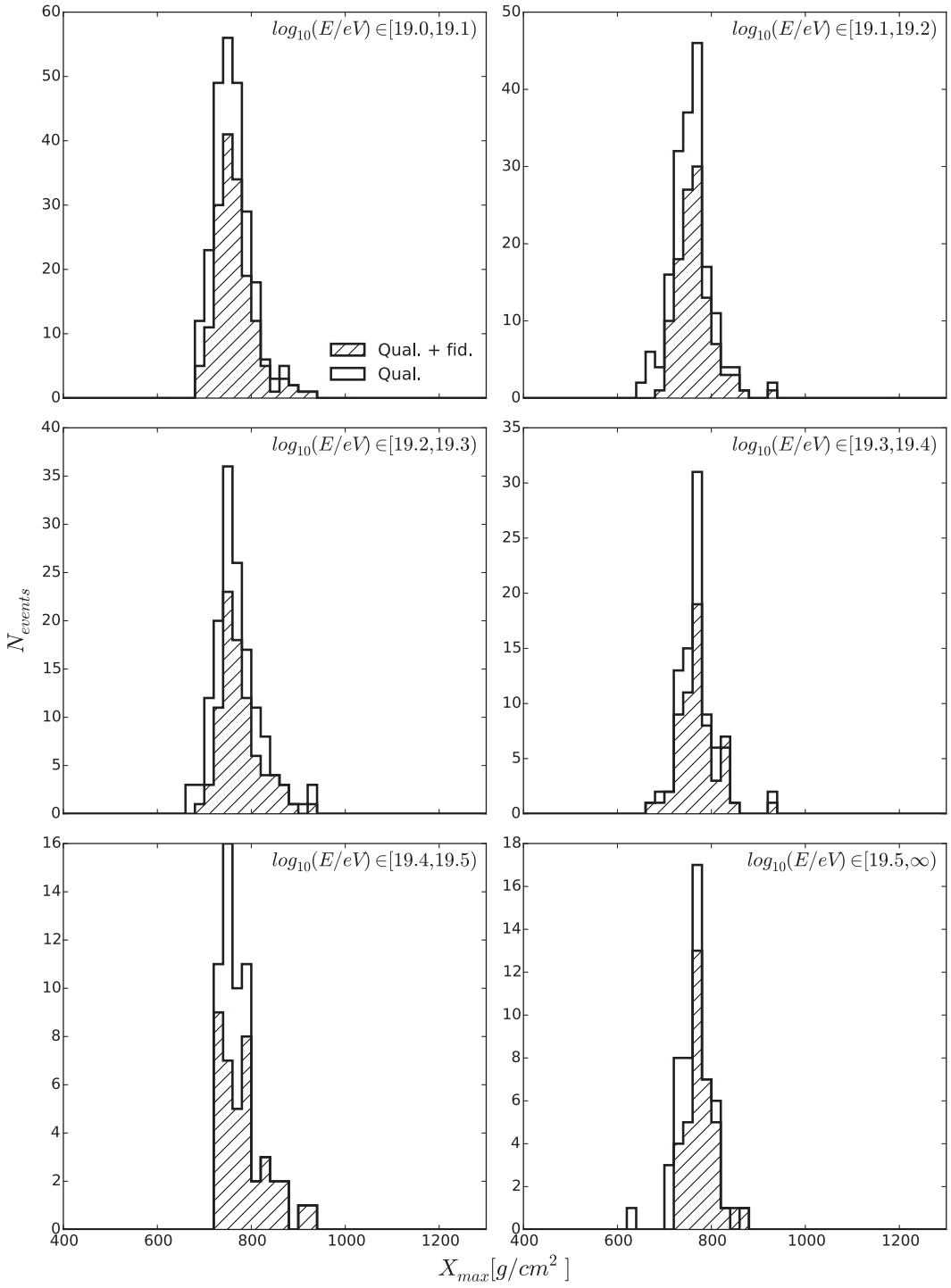


FIGURE 4.10: Same as FIGURE 4.8 for the highest energy ranges as labelled.

4.4 The $X_{\max}$ distributions

To study the $X_{\max}$ distributions we need to compare measurements with expectations. In the absence of an analytic expression for the theoretical $X_{\max}$ distributions we used for the composition analysis the approach given in [72] where the Monte Carlo simulations of several primaries using CONEX are fitted to a Generalised Gumbel distribution [73].

The chosen Generalised Gumbel distribution with location μ, scale σ and shape λ, has a probability density function given by:

$$f(x) = \frac{\lambda^\lambda}{\sigma\Gamma(\lambda)} e^{-\lambda\left[\frac{x-\mu}{\sigma} + e^{-\frac{x-\mu}{\sigma}}\right]}, \tag{4.26}$$

where $\Gamma(\lambda)$ is the Euler's Gamma function. We present the characterisation given in [72] of μ, σ and λ in terms of primary mass A and energy E:

$$\mu(A, E) = m_0 + m_1 \log_{10}(E/E_0) + m_2 \log_{10}^2(E/E_0), \tag{4.27}$$

$$\sigma(A, E) = s_0 + s_1 \log_{10}(E/E_0), \tag{4.28}$$

$$\lambda(A, E) = l_0 + l_1 \log_{10}(E/E_0), \tag{4.29}$$

where $E_0 = 10$ EeV is a reference energy. The dependence on the primary mass is given by m_i, s_i and l_i which are second order polynomials in $\ln(A)$ except m_2 which is a first order polynomial. The parameters of these polynomials are different for the different hadronic interaction models. A comparison of Monte Carlo simulations and Generalised Gumbel distributions for different primaries and hadronic models is shown in FIGURE 4.11 for completitude.

The moments of the parameterised $X_{\max}$ distribution (EQUATION 4.26) can be easily obtained through the derivations of the moment generating function:

$$M(t) = E[e^{tx}] = \lambda^{\sigma t} e^{\mu t} \frac{\Gamma(\lambda - \sigma t)}{\Gamma(\lambda)}, \tag{4.30}$$

$$\langle X_{\max} \rangle = \frac{\partial}{\partial t} M(t)\bigg|_{t=0} = \mu + \sigma \left[\ln(\lambda) - \psi_0(\lambda)\right], \tag{4.31}$$

$$\left\langle X_{\max}^2 \right\rangle = \frac{\partial^2}{\partial t^2} M(t)\bigg|_{t=0} = \sigma^2 \psi_1(\lambda) + \left\{\mu + \sigma \left[\ln(\lambda) - \psi_0(\lambda)\right]\right\}^2, \tag{4.32}$$

$$\sigma(X_{\max}) = \sigma\sqrt{\psi_1(\lambda)}, \tag{4.33}$$

where ψ_0 and ψ_1 are the polygamma[1] functions of order zero and one.

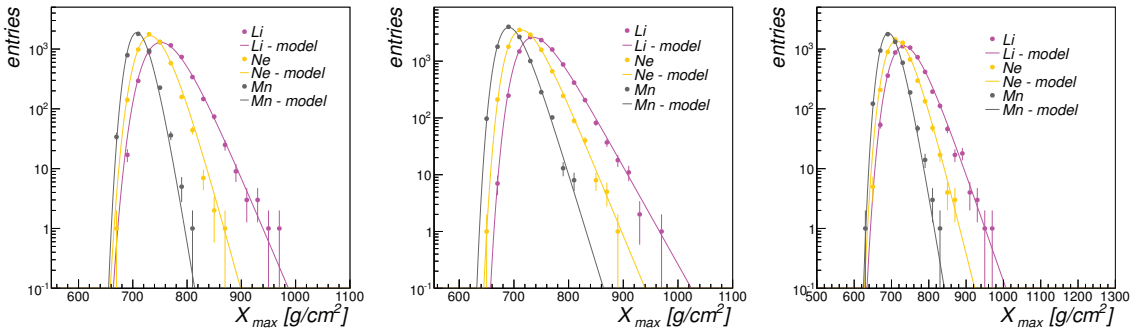


FIGURE 4.11: $X_{\max}$ distributions from simulations (points) at 10 EeV initiated by Li (pink), Ne (yellow) and Mn (grey) nuclei using EPOS LHC (left), SIBYLL 2.1 (middle) and QGSJETII-04 (right) hadronic interaction models. The modelled Gumbel distributions are represented by the continuous lines. Figure taken from [72].

Once the theoretical distributions, the efficiency and the response have been parameterised for a given model, the $X_{\max}$ distributions that would be observed by the detector can be evaluated. For a primary j the observed $X_{\max}$ distribution $g_j(X_{\max}|E)$ at a given energy E and expected field of view $[X_{\rm l}, X_{\rm u}]$, *i.e.* a given geometry, is related with the theoretical distribution $g_j(X^{true}_{\max}|E)$ (parameterised as a Generalised Gumbel distribution) through:

$$g_j(X_{\max}|E, X_{\rm l}, X_{\rm u}) = \frac{1}{\mathcal{N}}\chi(X_{\rm l}, X_{\rm u}) \int_{-\infty}^{\infty} \mathcal{R}(X_{\max} - X^{true}_{\max}|E)\epsilon(X^{true}_{\max}|E)g_j(X^{true}_{\max}|E)dX^{true}_{\max}, \tag{4.34}$$

where the response function $\mathcal{R}(X_{\max} - X^{true}_{\max}|E)$ and the efficiency $\epsilon(X^{true}_{\max}|E)$ are different for the two sets of events, those the anti-bias cut those without it as described in SECTIONS 4.2.1-4.2.2. The function χ is the characteristic function of the expected field of view (EQUATION 3.69). The normalisation constant is given by:

$$\mathcal{N} = \int_{X_{\rm l}}^{X_{\rm u}} dX_{\max} \int_{-\infty}^{\infty} \mathcal{R}(X_{\max} - X^{true}_{\max}|E)\epsilon(X^{true}_{\max}|E)g_j(X^{true}_{\max}|E)dX^{true}_{\max}. \tag{4.35}$$

Since the data distributions and moments are calculated on a bin-by-bin basis and the expected field of view changes for each event it is necessary to know the joint distribution of energy and expected field of view $f(E, X_{\rm l}, X_{\rm u})$. We assume that this

[1] $\psi_m(z) = \frac{d^m}{dz^m}\psi(z) = \frac{d^{m+1}}{dz^{m+1}}\ln\Gamma(z)$.

distribution is given by the data. In this way we obtain that the $X_{\max}$ distribution in a bin defined by the minimum energy E_{min} and maximum energy E_{max} is given by:

$$\begin{aligned} g_j(X_{\max}) &= \int_{E_{min}}^{E_{max}} dE \int_{\Omega(X_{\rm l})} dX_l \int_{\Omega(X_{\rm u})} dX_u g_j(X_{\max}|E, X_{\rm l}, X_{\rm u}) f(E, X_{\rm l}, X_{\rm u}) \\ &= E_{f(E,X_{\rm l},X_{\rm u})}[g_j(X_{\max}|E, X_{\rm l}, X_{\rm u})] \approx \frac{1}{N}\sum_{i=1}^{N} g_j(X_{\max}|E_i, X_{{\rm l},i}, X_{{\rm u},i}). \end{aligned} \tag{4.36}$$

Here, $\Omega(X_{\rm l})$ and $\Omega(X_{\rm u})$ symbolically denote the space of $X_{\rm l}$ and $X_{\rm u}$; N is the number of events in the bin. Equation 4.36 is just an identity. The integral $g_j(X_{\max}) = \int_{E_{min}}^{E_{max}} dE \int_{\Omega(X_{\rm l})} dX_l \int_{\Omega(X_{\rm u})} dX_u g_j(X_{\max}|E, X_{\rm l}, X_{\rm u}) f(E, X_{\rm l}, X_{\rm u})$ could be understood as the expected value of the function $g_j(X_{\max}|E, X_{\rm l}, X_{\rm u})$ where E, $X_{\rm l}$ and $X_{\rm u}$ follow the distribution $f(E, X_{\rm l}, X_{\rm u})$. That is why we denote it as $E_{f(E,X_{\rm l},X_{\rm u})}[g_j(X_{\max}|E, X_{\rm l}, X_{\rm u})]$. Finally, if we assume that $f(E, X_{\rm l}, X_{\rm u})$ is well characterised by the data, then we can approximate the integral by $\sum_{i=1}^{N} g_j(X_{\max}|E_i, X_{{\rm l},i}, X_{{\rm u},i})/N$ where N is the number of events in the energy bin.

When the $X_{\max}$ distribution is calculated using Equation 4.36 one can obtain the moments of the observed distributions for the different primaries using the usual expressions:

$$\begin{aligned} \langle X_{\max} \rangle_j &= \int X_{\max} g_j(X_{\max}) dX_{\max} \\ \sigma_j^2(X_{\max}) &= \int (X_{\max} - \langle X_{\max} \rangle_j)^2 g_j(X_{\max}) dX_{\max} \end{aligned} \tag{4.37}$$

The moments of Equation 4.37 are evaluated numerically. The comparison between the observed data moments and the theoretical predictions of the moments is displayed in Figure 4.12. Note that with "theoretical predictions of the moments" we mean the moments that would be observed by the detector if the composition was pure proton or pure iron.

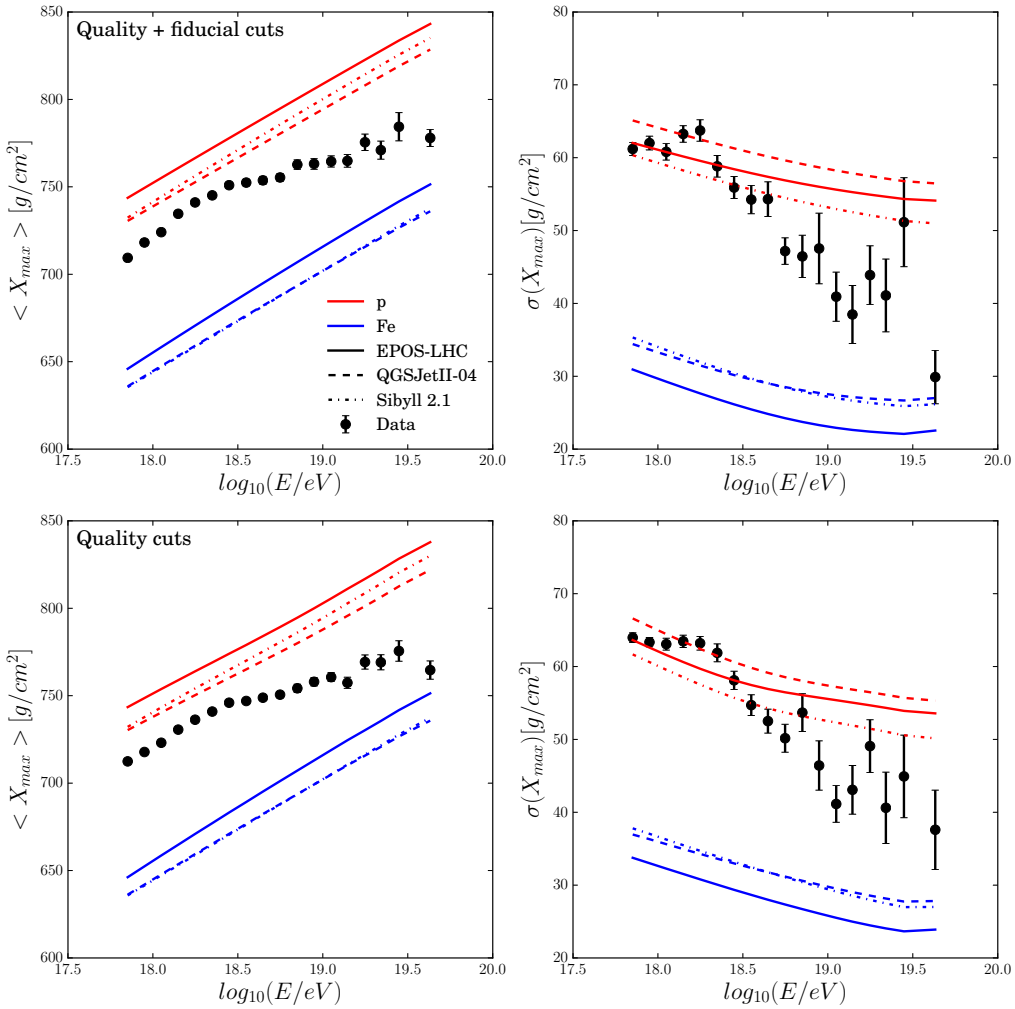


FIGURE 4.12: Mean value (left panels) and standard deviation (right panels) as a function of energy for the events applying the anti-bias cut (top panels) and without anti-bias cut (bottom panels). Data is represented using black circles. The proton (red) and iron (blue) predictions are calculated for three hadronic interaction models: EPOS LHC (continuous lines), QGSJETII-04 (dashed lines) and SIBYLL 2.1 (dash-dotted lines).

To compare the data moments with the theoretical predictions we could choose two approaches: we could subtract the detection effects from the observed data or we could apply these effects to the expected distributions. We choose the second approach. This is the procedure followed by the Telescope Array Collaboration in [74]. The theoretical predictions of the moments are different for the two data samples. Comparing the data moments relative to the expected moments of protons and irons, we conclude that a mixture composition is necessary to describe the mean and standard deviation at the same time. In addition, the data indicate a change of

composition at an energy on the range $\log_{10}(E/\text{eV})$ between 18.3 and 18.5 towards larger masses. These conclusions have been presented by the collaboration [67].

4.5 Algorithm for the Bayesian inference of the composition

One of the greatest disadvantages that the Bayesian analysis presents is the high level of computational requirements. We recall that the approach requires integrals over the space of the parameter of interest which could be multidimensional. In our case the parameter of interest is the relative composition of the cosmic rays and its dimension depends on the number of primary nuclei assumed.

Several numerical algorithms have been considered for the analysis of the $X_{\max}$ data. When the number of primaries is low one can perform a grid in the composition space and evaluate the posterior distribution at each point. We got good results using this method up to 4 primaries. This method, perhaps the most robust, is not practical when the dimension of the space increases.

The development of the Monte Carlo techniques provide new windows to perform Bayesian analyses. Several methods have been explored, for instance the Metropolis-Hastings algorithm, the Hamiltonian Markov Chain (see [75] for a review with applications of the last two methods) and the Affine Invariance Markov Chain Monte Carlo (see [76]). In all these methods the posterior distribution is first estimated. The evidence, *i.e.* its integral, must be calculated afterwards using approximations.

The final method chosen to infer the composition and to calculate the evidence in this work is the Skilling's nested sampling [77]. This method was originated to obtain in a simple way the Bayesian evidence in cases where it would be very difficult to obtain it because of the complexity of the likelihood function and/or the large number of dimensions of the integrals. This method allows us to estimate first the evidence and then the posterior distribution. As explained in Section 2.6.4 the evidence takes an important role in statistical inference because the Bayesian comparison between different models or scenarios is done by comparing their evidences. We recall that the evidence is defined as

$$Z = \int \mathcal{L}(\theta)\pi(\theta)d^k\theta \tag{4.38}$$

where we have omitted the dependence on data in the likelihood to simplify the notation; θ represents the parameter of interest (k-dimensional) and $\pi(\theta)$ is the assumed prior p.d.f.

The Skilling's method uses a change of variables that converts multidimensional integrals into a one-dimensional integral. This transformation is crucial for the method and we describe it in some detail. We define the "prior mass" $X(\lambda)$ as the amount of prior volume enclosed within the likelihood contour defined by $\mathcal{L}(\theta) = \lambda$, *i.e.* the θ space over which $\mathcal{L}(\theta) > \lambda$:

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} \pi(\theta) d^k\theta, \tag{4.39}$$

thus

$$dX = \pi(\theta) d^k\theta \tag{4.40}$$

The parameter λ takes values from 0 to $\mathcal{L}_{max}$ while the primary mass X ranges from 0 to 1 because is the integral of a probability density function. It is a monotonically decreasing function of λ as shown in FIGURE 4.13.

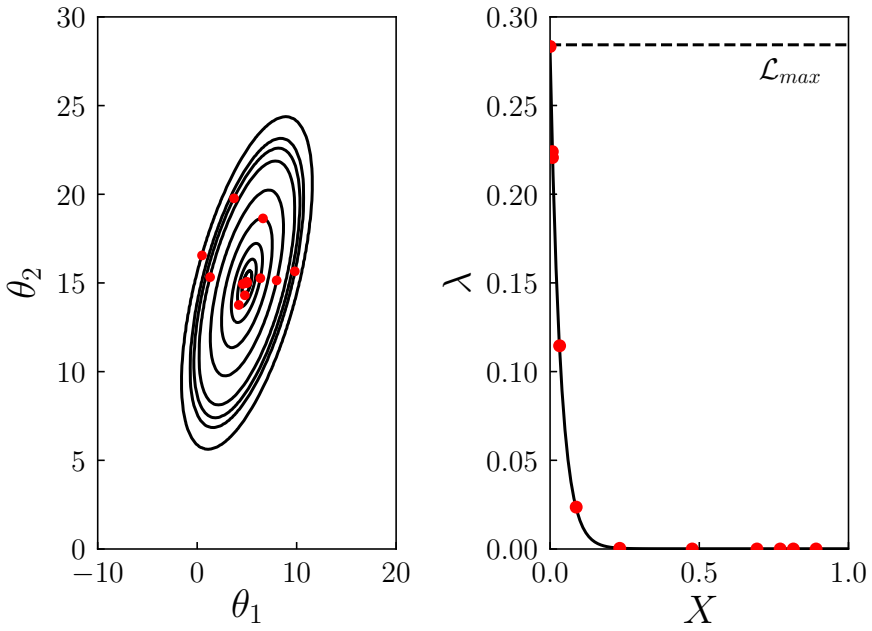


FIGURE 4.13: Example of a transformation from $\mathcal{L}(\theta)$ (left panel) to $\mathcal{L}(X)$ (right panel)

The evidence integral can be transformed using X into a one-dimensional integral:

$$Z = \int_0^1 \mathcal{L}(X) dX. \tag{4.41}$$

Numerical example

We are going to illustrate it with an example of a one-dimensional integral to clarify what is being done with the replacement of EQUATION 4.38 with EQUATION 4.41. Let θ be the parameter of interest which varies from 0 to 10 and let us assume that the likelihood of θ is given by a normal distribution (bounded in the the region $[0, 10]$) with mean $\langle\theta\rangle = 5$ and width $\sigma_\theta = 1$. We also assume that it has a prior distribution given by $\pi(\theta) = \text{Uniform}(0, 10) = 1/10$. Both the likelihood and the prior are plotted in FIGURE 4.14.

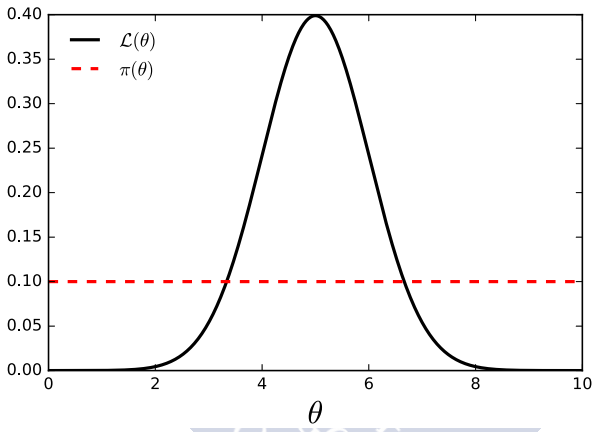


FIGURE 4.14: Likelihood function for the example (see text) as continuous black line and prior probability function as red dashed line.

The evidence in this example is $Z = \int \mathcal{L}(\theta)\pi(\theta)d\theta = 1/10$ (because we can neglect the integral of the Gaussian for $\theta < 0$ and $\theta > 10$).

For the numerical evaluation we select 1000 points from 0 to $\lambda_{max} = \mathcal{L}_{max}$ performing a partition in the likelihood domain instead of in the parameters domain. This procedure to integrate is similar to the Lebesgue integration. For each λ_i we calculate the prior volume $X_i = X(\lambda_i)$ (illustrated in FIGURE 4.15).

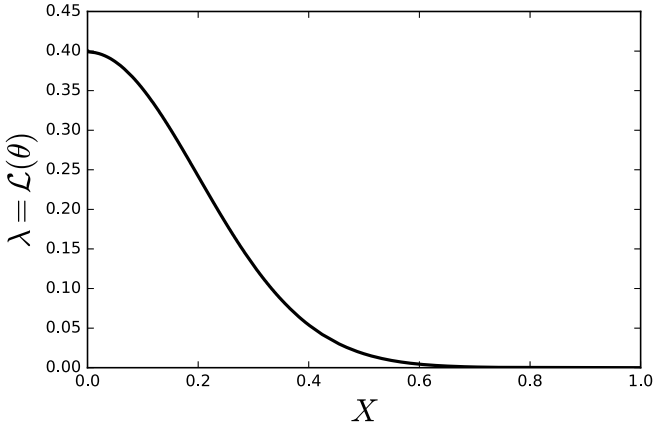


FIGURE 4.15: Likelihood of the example as a function of the prior volume.

The evaluation of the integral EQUATION 4.41 can be easily done as:

$$Z \approx \sum_k \omega_k \mathcal{L}_k, \tag{4.42}$$

where $\omega_k = X_{k-1} - X_k$ or $\omega_k = (X_{i-1} - X_{i+1})/2$ (trapezoidal rule). We use the former one.

Notice that this procedure is just to calculate the integral by using a simple numerical method. The advantage is that the integral in λ-X is always a one-dimensional integral. By using this numerical approach we obtain an evidence which differs from the exact evidence in 0.003%.

Nested sampling method

We are now going to explain the nested sampling method for the computation of the Bayesian evidence. It is an algorithm to select the points for the numerical integration and to obtain the evidence and its precision. The algorithm starts with N "active points" randomly generated from the prior distribution. Now, for each iteration we select the point with largest prior volume X^* (and hence, lowest likelihood, $\mathcal{L}^*$) and we discard this point from the list of "active points". After that we have $N-1$ points. Now a new point is randomly generated following the prior distribution and it is only accepted if $\mathcal{L}(\theta_{new}) > \mathcal{L}^*$, recovering in this way N points inside the domain bounded by X^* which is "nested" inside the old domain.. The lowest likelihood of the new list of active points has a value that is larger than that of the removed point (lowest likelihood of the previous active points). As a result the corresponding larges value

of X in the new list of active points is smaller than that obtained in the previous list. The procedure can be iterated until some stopping criterion is satisfied.

A graphic example of the procedure is shown in FIGURE 4.16.

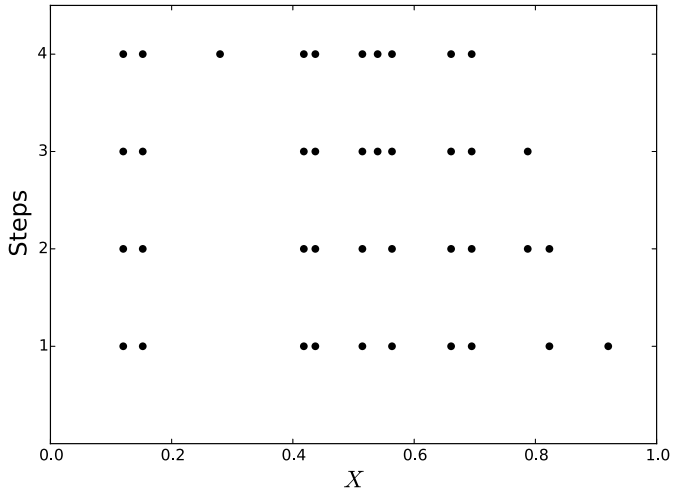


FIGURE 4.16: Example of the nested sampling method with 10 active points. Only four iterations are shown with 10 active points.

At each iteration i the "shrinkage ratio" $t_i = X_i^*/X_{i-1}^*$, where X_i^* is the prior mass of the point that has the lowest likelihood at iteration i . The shrinkage ratio is distributed as a Beta$(N, 1)$ which is the distribution of the N^{th} order statistic of an uniform distribution defined in the range $[0, 1]$. This distribution has an expectation $E[\ln(t)] = -1/N$ and standard deviation $\sigma(\ln(t)) = 1/N$. Since the shrinkage ratio at each iteration is independent of the other iterations, after i iterations the prior volume is expected to shrink down such that

$$\ln(X_i) \approx -(i \pm \sqrt{i})/N. \tag{4.43}$$

Thus, the expected value of X_i after i iterations is $\langle X_i \rangle = \exp(-i/N)$. Notice that by using these properties we can avoid the evaluation of the different X_i after enough iterations. Then we can calculate the evidence as EQUATION 4.42.

Note that as we increase the number of iterations we are shifting the N active points to smaller X values (as shown in FIGURE 4.16) and we have calculated the contribution to the integral in the X range of the points that we have being discarding using EQUATION 4.42. As a result when the iteration process stops there is a missing contribution due to the X range that is covered by the latest set of N active points

which has no been explored. We can refine the evaluation of the inference by estimating the integral of the final set of N active points (assuming that the likelihood is constant is this region) as:

$$Z = Z_l + \sum_{j=1}^{N} \omega^* \mathcal{L}_j. \tag{4.44}$$

Here, ω^* is the weight of the last iteration of the algorithm and Z_l is the estimated evidence using the discarded points. This last increment ought to be unimportant because there should have been sufficient iterations to accumulate most of the integral. The nested sampling algorithm is summarised below:

Algorithm 1 Skilling's nested sampling for Bayesian evidence

1: Generate N "active points" from the prior $\pi(\theta)$
2: Set $Z = 0$ and $X_0 = 1$.
3: Set $i = 1$
4: **while** not stopping criterion satisfied **do**
5: Set $(\mathcal{L}^*, X^*)$ the point with lowest likelihood and largest X.
6: Set $X_i = \exp(\frac{-i}{N})$ or sample $t_i \sim p(t) = Nt^{N-1}$ and set $X_i = t_i X_{i-1}$
7: Set $\omega_i = X_{i-1} - X_i$ or $\omega_i = (X_{i-1} + X_{i+1})/2$
8: $Z \leftarrow Z + \mathcal{L}^* \omega_i$
9: Generate $\theta_{new}/\mathcal{L}(\theta_{new}) > \mathcal{L}^*$ following the prior $\pi(\theta)$
10: $Z \leftarrow Z + \frac{X^*}{N}(\sum_{j=1}^{N} \mathcal{L}(\theta_j))$

The posterior inferences can be easily obtained using all generated points during the iterations of the nested sampling. For each point θ_i we can assign its probability as:

$$p_i = \frac{\omega_i \mathcal{L}_i}{Z}. \tag{4.45}$$

Then, we can construct the posterior probability density function as a table $p_i(\theta_i)$. It is then straightforward to calculate the mean of the posterior probability density function, the marginals, etc.

Sampling within the constrained prior volume plays an important role for the algorithm. Notice that the actual sampling method used to select a new θ point is not important for the calculation. However it can make all the difference from the computational point of view. The described method becomes inefficient as the prior volume is reduced. In his paper ([77]) Skilling proposes to perform a Markov Chain

Monte Carlo in order to sample new points (a good review about the application of these methods can be found in [75]).

An important contribution to an efficient sampling of the new point at each iteration was presented in [78]. Here the authors proposed to fit an ellipsoid bounding the N active points using the covariance of the active points and then to sample points within this ellipsoid (multiplied by some enlargement factor, typically 1.2). This method is computationally efficient and robust if the posterior probability density function is unimodal, as it is shown in [79]. For distributions that are not unimodal the authors of [79] improved the ellipsoidal nested sampling by introducing a partition of the active points using a "$k-$means clustering algorithm" (see [80]) and fitting each partition with ellipsoids. This algorithm is the so called MultiNest algorithm and it is also able to give good results when the posterior probability density function presents more than one mode. For the partitioning the active points into clusters the MultiNest algorithm requires the points to be uniformly distributed in an unit hypercube (see [79] for the details). In Appendix B we show how we transform the unit hypercube into the composition space. For the analysis of the $X_{\max}$ data we will present the results using the MultiNest algorithm but the analysis was performed also using the ellipsoidal method without significant differences neither in the posterior distribution nor in the evidence.

4.5.1 Stopping criterion and uncertainty of the evidence

Notice that Skilling's algorithm does not specify the stopping criterion. For the stopping criterion we use an estimation of the remaining evidence, *i.e.* what is added at each iteration. An approximation of the remaining evidence is given using the highest likelihood of the current set of active points. At iteration j we assume that the remaining evidence is $\hat{Z} = X_j \mathcal{L}_{max}$, where $\mathcal{L}_{max}$ is the maximum value of likelihood of the current set of active points. If the calculated evidence at this iteration is Z_j, then, the stopping criterion is based on the estimation of the remaining increment of the evidence:

$$\ln\left(\frac{\hat{Z} + Z_j}{Z_j}\right) < \beta \tag{4.46}$$

where β is some threshold. When the increment in the calculated evidence is estimated to be below this threshold we stop the nested sampling. For the analysis of

the composition we are going to use $\beta = 0.5$. For testing porpoises we have used different values of β without finding significant differences.

The calculation of the uncertainty is more complicated. We are going to show how one of the most important variables in the Bayesian analysis (the *information*, H) is related with the uncertainty. The information is defined as the negative Shannon's entropy (see [81]). For a probability density function $p(x)$, the information is given by:

$$H = -S = \int p(x)\ln(p(x))dx. \tag{4.47}$$

The information is measured in *natural units of information* (*nat*)[2]. In Bayesian statistics the most common quantity used related with the information is the *relative information* of the posterior with respect to the prior, $\mathcal{H}$. The relative information is mathematically expressed as the Kullback-Leibler divergence ([82]) from the prior to the posterior and represents the information gain if the posterior is used instead of the prior or, in other words, the information gained when the data is analysed (moving from the prior to the posterior). We assume that we want to infer a certain parameter θ and we measure some data set D. Let $\pi(\theta|D)$ and $\pi(\theta)$ be the posterior and prior probability density functions respectively. The relative information is given by:

$$\mathcal{H} = D_{KL}(\pi(\theta|D)||\pi(\theta)) = \int \pi(\theta|D)\ln\left[\frac{\pi(\theta|D)}{\pi(\theta)}\right]d\theta \tag{4.48}$$

For our numerical approach it can be calculated in the following way: if the loop of the nested sampling is stopped at iteration k, the relative entropy is given by:

$$\mathcal{H} \approx \sum_i \frac{\omega_k \mathcal{L}_k}{Z}\ln\left[\frac{\mathcal{L}_k}{Z}\right]. \tag{4.49}$$

The actual numerical uncertainty is given by $Z_{est} - Z_{true}$ but we of course cannot make this comparison because we do not know Z_{true} when we analyse actual data. One method to estimate the numerical uncertainty is performing the analysis several times and by observing the distribution of the results. Nevertheless, we are going to try to find a method which gives us an estimate of the uncertainty by performing only one trial.

At each iteration, the prior volume is reduced to regions of larger likelihoods. The main contribution to the uncertainty of the evidence calculation comes from the

[2]EQUATION 4.47 is sometimes written changing the natural logarithm by a binary logarithm and then the information is measured in *bits*.

sampling of X. In particular, in sight of EQUATION 4.46 we have that the main uncertainty is given by the value of X at the last iteration because its likelihood is maximal. Let m be the last iteration and N the number of active points selected for the nested sampling algorithm. Since $X_m = tX_{m-1}$ where t follows a probability function $P(t) = Nt^{N-1}$, then we will arrive to EQUATION 4.43. It follows that the uncertainty in the evidence evaluation is then related with the uncertainty in X_m, the volume in which the N active points are confined:

$$\delta(\ln Z) \sim \delta(\ln \omega_m) \sim \delta(\ln X_m) = \frac{\sqrt{m}}{N}. \tag{4.50}$$

We are now going to relate this uncertainty with the relative information $\mathcal{H}$. $\mathcal{H}$ measures the information gain from the prior to posterior, *i.e.*, how "peaked" the likelihood is. If the likelihood is peaked with respect to the prior near $X = 0$, the numerical integration will introduce errors. On the other hand, if the likelihood is not so much peaked with respect to the prior, the approximation in the last iteration describes well the non-explored area.

As an example assume that the likelihood function is zero for all X except for the last iteration and has a constant value $\mathcal{L}(X) = \mathcal{L}_{max}$ for all $X < X_m$. For this example we have that the evidence is $Z = \mathcal{L}_{max} X_m$ and the relative information is just:

$$\mathcal{H} = \int_0^1 P(X)\ln(P(X))dX = \int_0^{X_m} \frac{1}{X_m} \ln\left(\frac{1}{X_m}\right) dX = -\ln X_m. \tag{4.51}$$

Replacing the result of EQUATION 4.51 into EQUATION 4.43 we obtain that the most relevant region for the evaluation of the evidence is at $m = \mathcal{H}N$. The uncertainty can be then expressed as:

$$\delta(\ln Z) \sim \delta(\ln X_m) = \sqrt{\frac{\mathcal{H}}{N}}. \tag{4.52}$$

In sight of this example we can assume that the dominant uncertainty in the calculation of the evidence is given by $\mathcal{H}/N$, or, at least, we can assume that the relative information is related with the numerical uncertainty of the evidence.

As a computational check of the algorithm we perform 1000 trials applying the nested sampling algorithm to obtain the evidence and its uncertainty in a case where the exact value of the integral and the relative information can be calculated analytically. We use again the example used in SECTION 4.5, where $\mathcal{L}(\theta) = \mathcal{N}(5, 1)$ and $\pi(y) =$

Uniform$(0, 10)$. The evidence and relative information are:

$$Z = \int_0^{10} \mathcal{L}(y)\pi(y)dy = \frac{1}{10}, \tag{4.53}$$

$$\mathcal{H} = \int_0^{10} \frac{\mathcal{L}(y)\pi(y)}{Z} \ln\left(\frac{\mathcal{L}(y)}{Z}\right) dy = -\ln(\sqrt{2\pi e}) + \ln(10) \approx 0.884. \tag{4.54}$$

According to the previous discussion, since we now use 3000 active points for the nested sampling algorithm, we can expect to get a numerical uncertainty in the evaluation of the evidence of order $\sqrt{\mathcal{H}/N} \approx 0.017$. In FIGURE 4.17 we display the difference between the estimated value Z_{est} in each trial and the exact value Z_{true}. The mean value of $\ln Z_{est} - \ln Z_{true}$ is 1.3×10^{-4} with a standard deviation equal to 0.018. On the other hand, the estimated uncertainty of the evaluation of the integral by using EQUATION 4.52 is also displayed in the same figure for each trial, obtaining a mean value $\langle \delta(\ln Z_{est}) \rangle = 0.017$. After this computational study we can conclude that both evidence and relative information are well estimated and that the relative information (their square root divided by the number of active points) gives us a good idea about the numerical uncertainty of algorithm in the evaluation of the evidence for one single run and it is not necessary to perform several trials for each analysis.

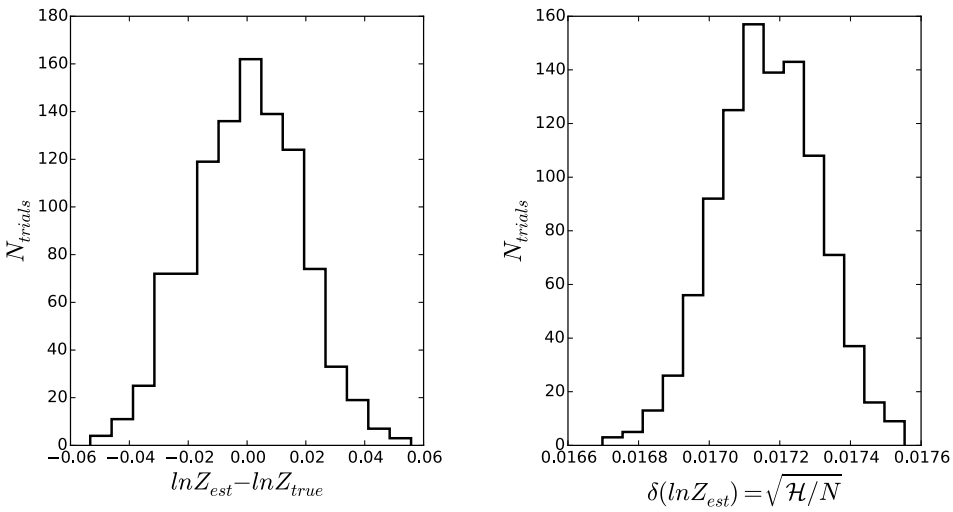


FIGURE 4.17: Left: histogram of $\ln Z_{est} - \ln Z_{true}$ for the 1000 performed trials for testing the numerical uncertainty of the algorithm (see text). Right: histogram of the estimated numerical uncertainty.

4.6 The prior predictive distributions

Comparing the prior predictive distributions with the observed data could be interesting before the analysis. This comparison can give us an idea about our prior assumptions. It is a sort of average of all possible distributions. For this analysis we are only going to assume that the sum of fractions of the different primaries is one. For example, if we assume a proton-iron scenario, then the fraction of protons plus the fraction of irons is one. In a four primary scenario proton-helium-nitrogen-iron, the sum of the four fractions is one and all possible combinations of the fractions such that the sum of the fractions is one is considered equally likely. These conditions describes the so called "flat prior". An example of a flat prior is shown in FIGURE 4.18 where a p-He-N-Fe scenario is considered.

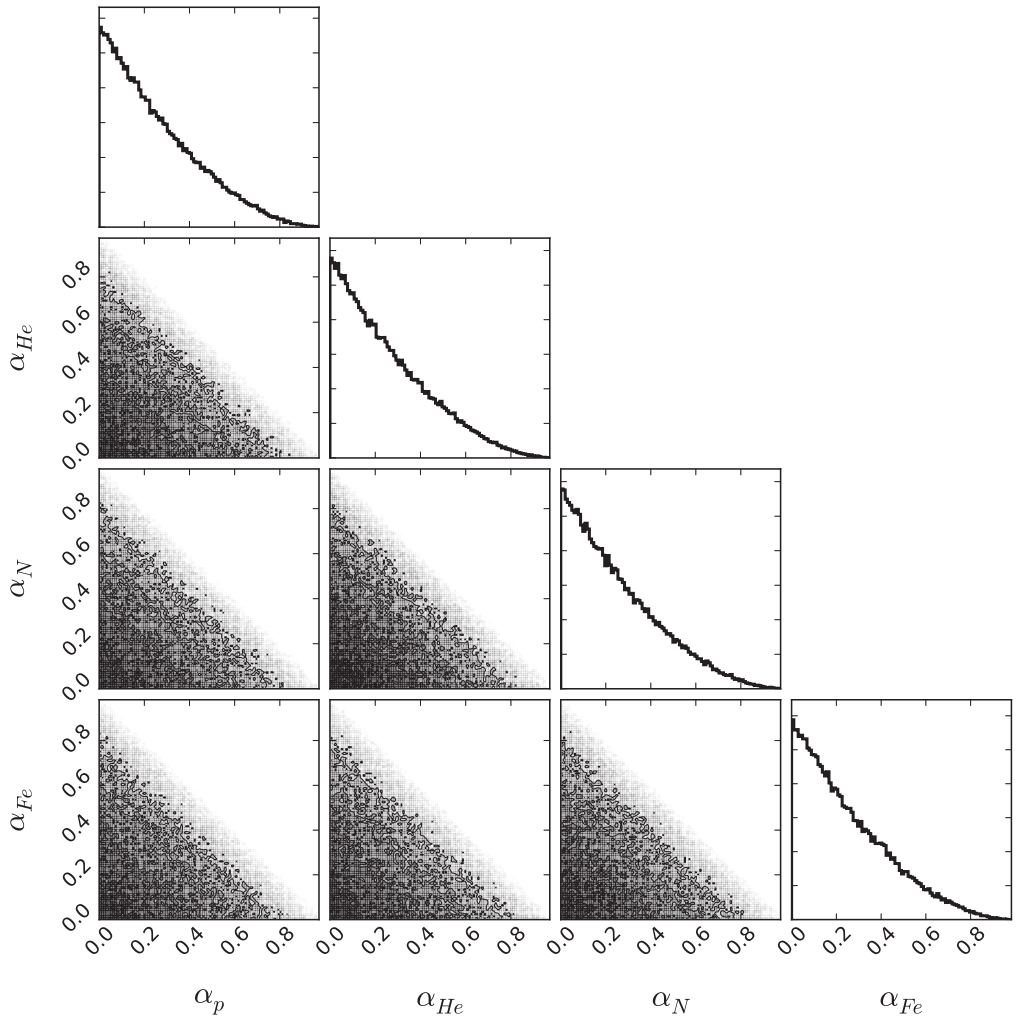


FIGURE 4.18: A sample of 10^5 points following the prior distribution in the composition. The points are "flat" distributed in a 4D space.

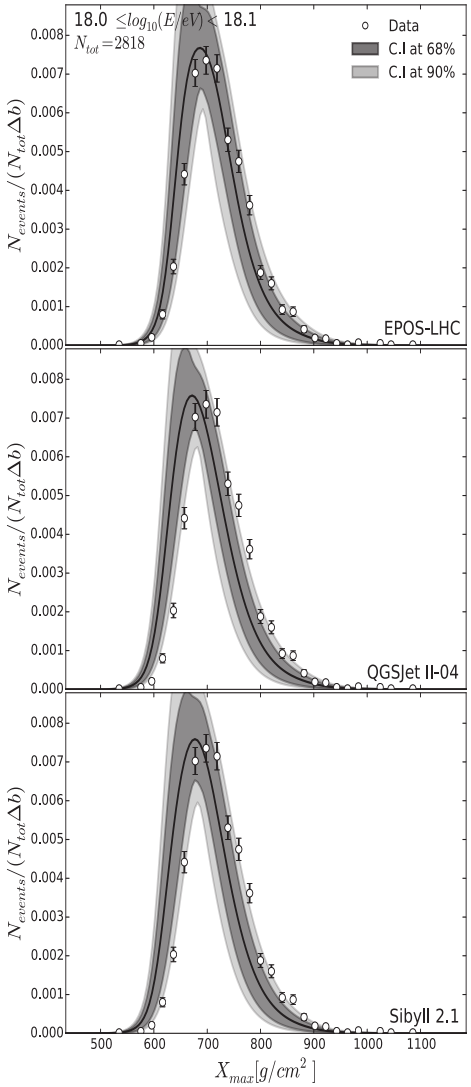


FIGURE 4.19: Prior predictive distribution for using a flat prior (grey bands) compared the observed data distribution (white points) at energy bin $18 \leq \log_{10}(E/\text{eV}) < 18.1$.

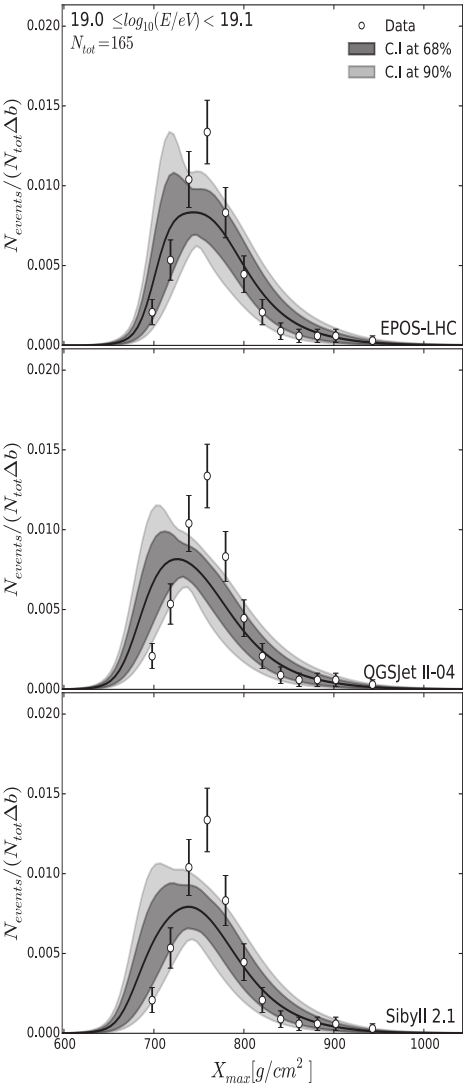


FIGURE 4.20: Prior predictive distribution for using a flat prior (grey bands) compared the observed data distribution (white points) at energy bin $19 \leq \log_{10}(E/\text{eV}) < 19.1$.

The prior predictive distributions are shown in FIGURES 4.19-4.20 for two different energy bins. In this case, the black line gives a relative fraction of 0.25 for all primaries. We can say something more than just using the moments. For example: the actual composition is lighter than that given by the flat prior. We need more

p-He in the energy bin $18 \leq \log_{10}(E/\text{eV}) < 18.1$ to fit the right tail of the $X_{\max}$ distribution; in the energy bin $19 \leq \log_{10}(E/\text{eV}) < 19.1$ the lighter elements seem to fit quite well the data distribution but the iron seems overestimated. Notice that these conclusions are drawn from visual inspection and we must do a complete analysis to achieve valid conclusions. This is done in full detail in the next chapter.

4.7 Dealing with systematic uncertainties

To finalise this chapter we are going to explain formally how to deal with systematic uncertainties in a Bayesian approach and how we are going to deal with them in this work. Unfortunately, we cannot follow the formal procedure because we do not know the correlation between the parameters whose uncertainties act as systematic uncertainties in the composition analysis. To be clear we are going to present the mathematical treatment of the systematic uncertainties through an example but the reasoning can be extrapolable to a general case.

Assume again a simple example where the $X_{\max}$ data distribution is composed by protons and irons and the only "smearing" effect over the data is due to a Gaussian resolution described by $\mathcal{R}(X_{max} - X_{max}^{true}) = \mathcal{N}(\mu, \sigma)$. Since in this example the response function depends on two parameters, we make this dependence explicit:

$$\mathcal{R}(X_{max} - X_{max}^{true}) \rightarrow \mathcal{R}(X_{max} - X_{max}^{true}, \mu, \sigma) \tag{4.55}$$

Assume also in this hypothetical example that after a lot of simulations we can assure that the mean value of the response function is $\mu = 0$ with a 100% of confidence level but the width of the response has still certain uncertainty given by the distribution represented in FIGURE 4.21. The response function changes again:

$$\mathcal{R}(X_{max} - X_{max}^{true}, \mu, \sigma) \rightarrow \mathcal{R}(X_{max} - X_{max}^{true}, \sigma). \tag{4.56}$$

Here, we omit the dependence in μ because it takes always the value zero. The bivariate distribution $\mathcal{R}(X_{max} - X_{max}^{true}, \sigma)$ is shown if FIGURE 4.22. The uncertainty in the parameter σ acts as a systematic uncertainty in the composition analysis. Now we are going to proceed with the composition analysis reformulating the equations for this hypothetical example.

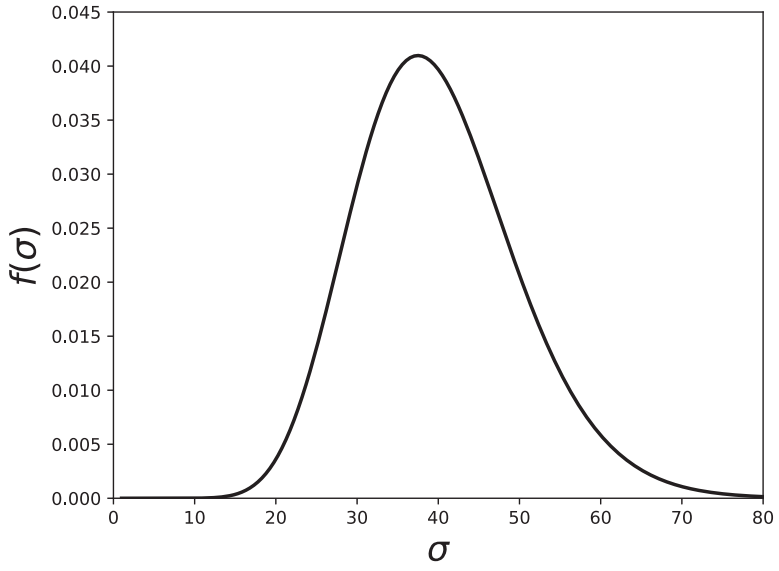


FIGURE 4.21: p.d.f of the width of the response function.

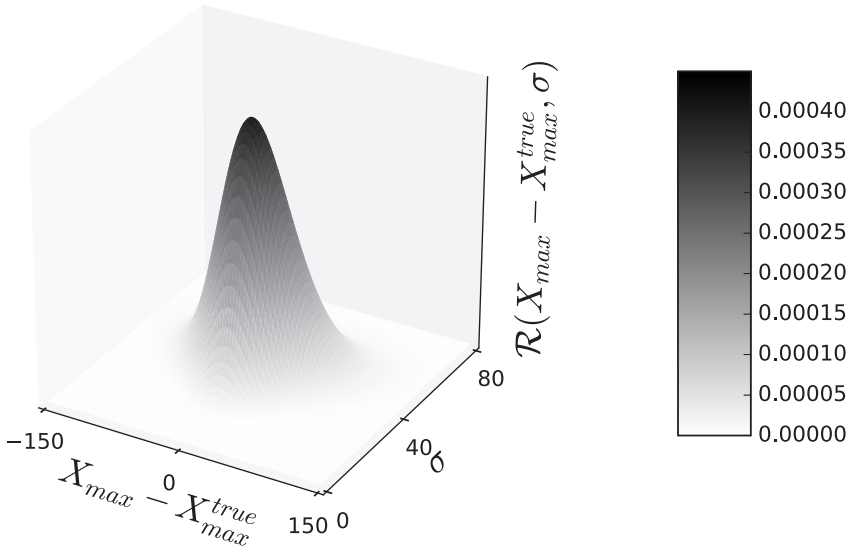


FIGURE 4.22: Joint p.d.f $\mathcal{R}(X_{max} - X_{max}^{true}, \sigma)$

Let $g_p(X_{\max}^{true})$ and $g_{Fe}(X_{\max}^{true})$ be the theoretical $X_{\max}$ distributions for protons and irons respectively. The actual data (*i.e*, without detector distortions) is given by

$$g(X_{\max}^{true}) = \alpha g_p(X_{\max}^{true}) + (1-\alpha) g_{Fe}(X_{\max}^{true}), \tag{4.57}$$

where the fraction of protons in data is represented by the parameter α. Nevertheless, the data recorded by the detector is given by

$$g(X_{\max}|\hat{\sigma}) = \int \mathcal{R}(X_{\max} - X^{true}_{\max}|\hat{\sigma}) \left[\alpha g_p(X^{true}_{\max}) + (1-\alpha) g_{Fe}(X^{true}_{\max})\right] dX^{true}_{\max}, \quad (4.58)$$

where $\hat{\sigma}$ is the true value of the width of the response function. Since we do not know the true value of σ we must perform our analysis taking into account this uncertainty.

The correct analysis is performed using the the $f(\sigma)$ distribution which represents our knowledge about the width of the response function. The posterior distribution of the composition fraction for all the possible widths is obtained by integrating over all widths:

$$\begin{aligned}
\pi(\alpha|D) &= \int \pi(\alpha|D,\sigma) f(\sigma) d\sigma = \\
&= \frac{\alpha}{Z} \int\int \mathcal{R}(X_{\max} - X^{true}_{\max}|\sigma) f(\sigma) g_p(X^{true}_{\max}) dX^{true}_{\max} d\sigma + \\
&+ \frac{1-\alpha}{Z} \int\int \mathcal{R}(X_{\max} - X^{true}_{\max}|\sigma) f(\sigma) g_{Fe}(X^{true}_{\max}) dX^{true}_{\max} d\sigma = \\
&= \frac{1}{Z} \int \mathcal{R}(X_{\max} - X^{true}_{\max}) \left[\alpha g_p(X^{true}_{\max}) + (1-\alpha) g_{Fe}(X^{true}_{\max})\right] dX^{true}_{\max}
\end{aligned} \quad (4.59)$$

Here, $\mathcal{R}(X_{\max} - X^{true}_{\max}) = \int \mathcal{R}(X_{\max} - X^{true}_{\max}|\sigma) f(\sigma) d\sigma$ and it is interpreted as the response function for all possible values of the parameter σ. The posterior distribution calculated using EQUATION 4.59 takes into account all sources of uncertainty: the statistical uncertainty given the data set D and the systematic uncertainty in the response function given by $f(\sigma)$.

As it has been said at the beginning of this section, we cannot follow the correct treatment of the systematics because we do not know neither the distributions nor the correlation of all variables whose uncertainties act as systematic uncertainties in the composition analysis. For this reason the systematic uncertainties presented in the next chapter, where we are going to analyse actual data, have been obtained by varying all the parameters of the response and efficiency functions combining the mean values and the extremes of this functions (maximum and minimum efficiency and maximum and minimum resolution) assuming that the systematic uncertainties in the composition are those that have the maximum and the minimum primary mass. We could proceed assuming the systematics by the extremes of each primary but then we would lose the correlation between the posterior fractions.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 5

$X_{\max}$ composition

In this chapter we present the results of the composition analysis of the $X_{\max}$ data using the Bayesian methods, the detector description and the data presented in the previous chapter. For the composition analysis we are going to assume that the cosmic rays arriving to the Earth can be a combination of protons (the lightest and more abundant nucleus of the Universe), iron nuclei (the most stable), helium and nitrogen nuclei[1]. These elements are approximately equispaced in the logarithm of their masses. We into account all possible combinations of these elements, assuming scenarios with only two components, three and four components. In addition, we consider one more scenario with six components by adding to the mentioned four the possible presence of lithium and silicon. Notice that all the other scenarios are subsets of this last case. Due to the limited number of events the assumption of the number of primaries in data can lead to different conclusions. Bayesian statistics allows a us to study the ability of current measurements to discriminate among the different primaries. Besides the different composition scenarios, we are going to assume three different high-energy hadronic interaction models: Epos LHC, QGSJetII-04 and Sibyll 2.1. Therefore, a total of 36 different scenarios are going to be analysed and compared with the aim of extracting the maximum information possible about the composition of the cosmic rays using $X_{\max}$ measured with the Pierre Auger Observatory. Moreover, the analysis is going to be performed using two data samples: one with fiducial cuts applied and another one without them. Since the statistical approximation, the algorithm and the differences between the data sets have been already discussed in the previous chapters, only the results are presented here.

[1] Along the text we can refer hydrogen nuclei as protons, helium nuclei as heliums, etc.

5.1 p-Fe scenario

We start with the p-Fe scenario. This contains the lightest and heaviest primaries considered in this work. This scenario is often considered. Since the different hadronic models have dissimilar distributions one should expect to infer different compositions when analysing with different hadronic models. This can be seen in FIGURE 5.1. It is remarkable that the composition obtained using the anti-bias cuts and without these cuts is almost identical but when the anti-bias cut is removed the number of events increases significantly and we obtain better estimations of the fractions[2]. One can observe some similarities and differences among the three hadronic models. The estimated composition is different. We have heavier composition for EPOS LHC along all the range of energy than the one obtained with the other models and QGSJETII-04 gives the lightest composition which has a proton fraction above 80% (with a 90% of confidence level) up to 10 EeV. The similarities can be seen in the trend of the composition change with energy. For the three models the composition of cosmic rays is mainly protons at $E = 10^{17.8}$eV. The proton fraction increases up to energies around $10^{18.4} - 10^{18.5}$ eV. At this point the proton fraction drops reaching a local minimum at the energy range $18.7 \leq \log_{10}(E/\text{eV}) < 18.8$. At higher energies, the composition of the cosmic rays becomes heavier.

The observed structure could be due to statistical fluctuations in the data but when the anti-bias cut is not applied the number of events increases and the same structure is observed. This fact strengthens our conclusions: the cosmic-ray composition becomes lighter from values of $\log_{10}(E/\text{eV})$ 17.8 up to 18.4. In the energy region $18.4 \leq \log_{10}(E/\text{eV}) < 18.8$ (or 10^{19} eV) the proton fraction first drops to a local minimum and then rises again to a local maximum. Above this energy the proton fraction drops again: the composition becomes heavier as the energy increases.

[2]In this context "better estimations" means estimations with less uncertainty.

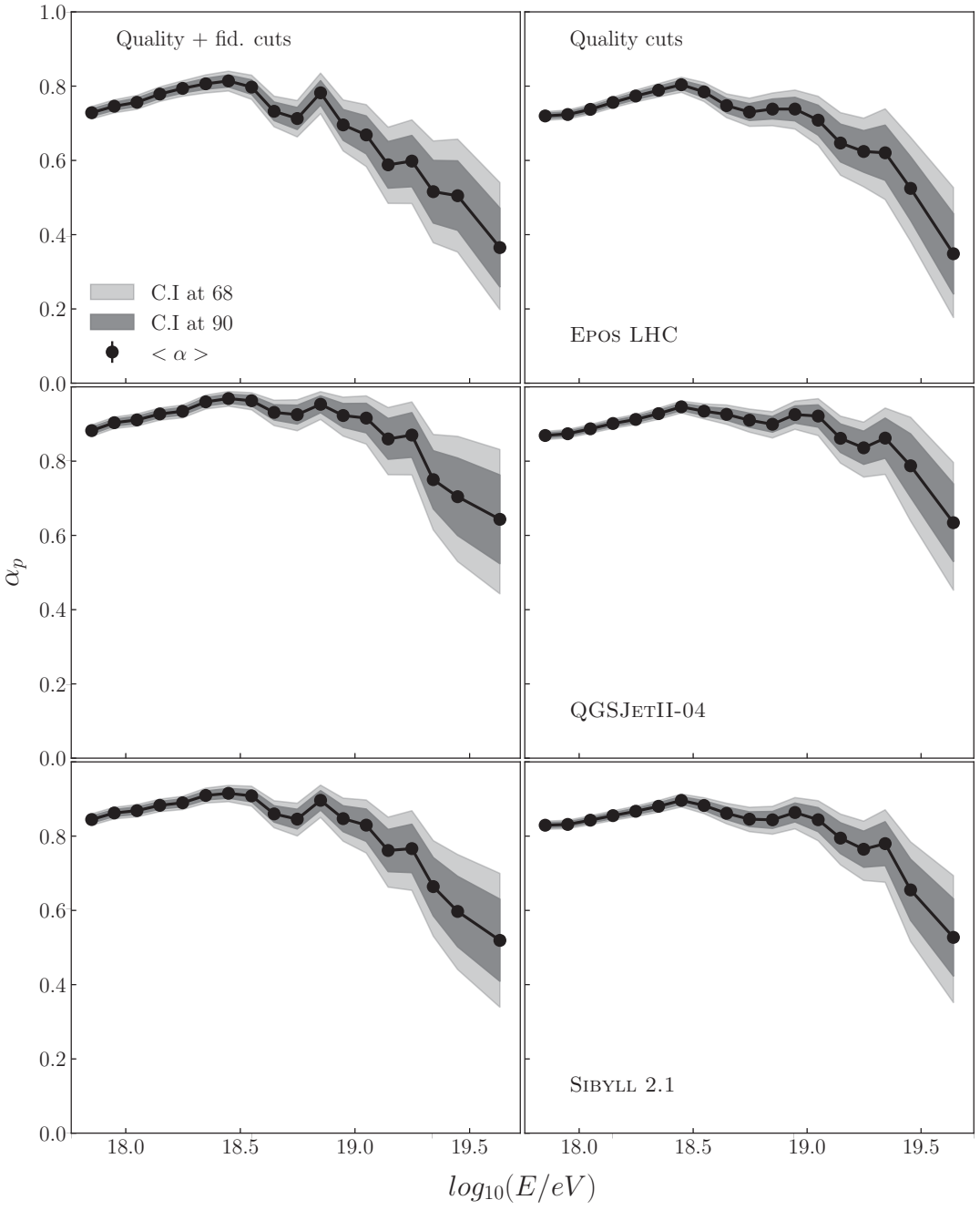
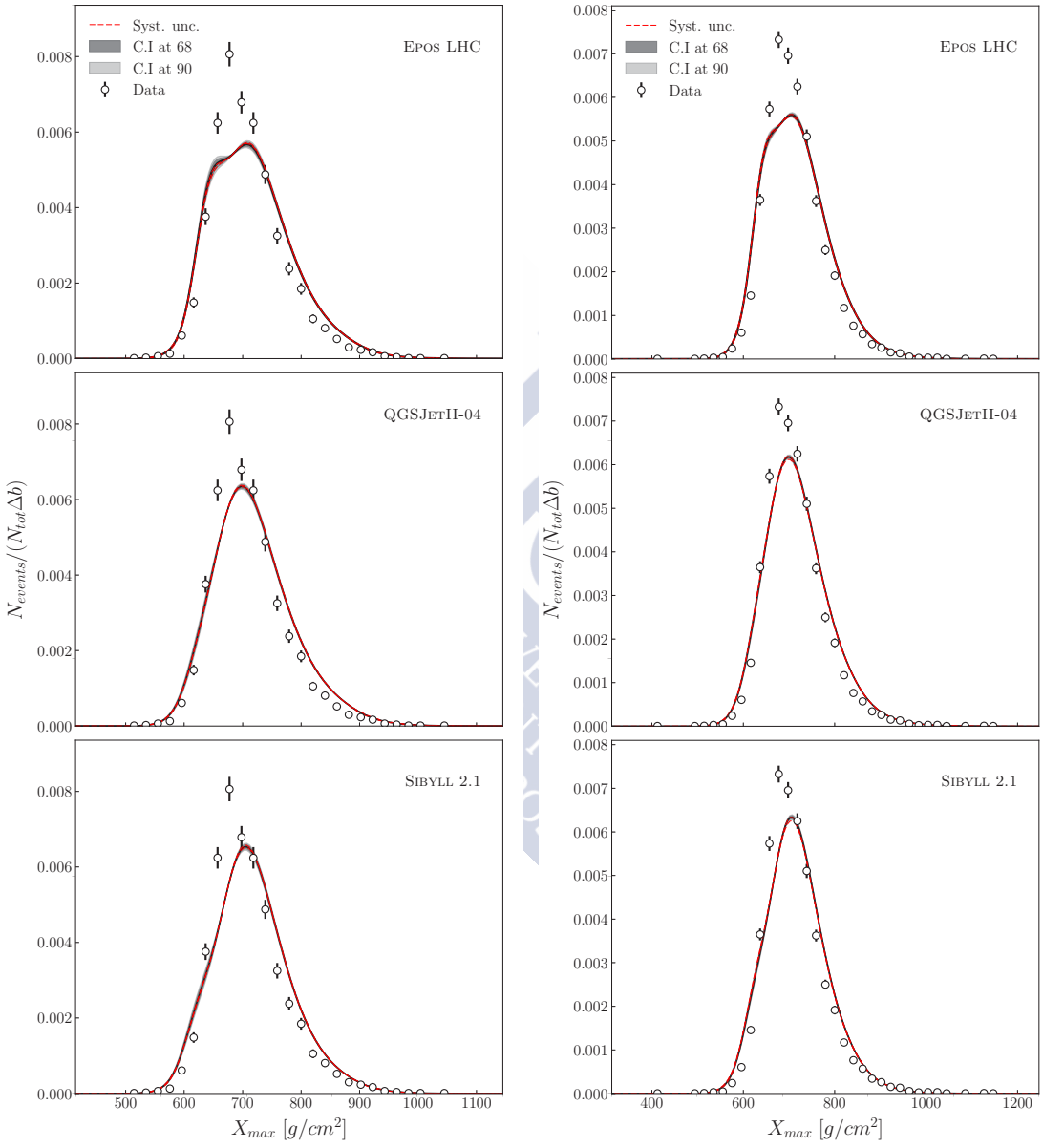


Figure 5.1: Trends of the proton fraction with the energy in the p-Fe scenario for the three hadronic models and using both data samples: with fiducial and without fiducial cuts. The shaded bands represent the confidence interval of the fraction at 90% (clearer) and 68% (darker). The mean value of the posterior p.d.f with the systematic uncertainties are shown as black circles with bars.

The question is if the p-Fe scenario is sufficient to describe the observed data. To

answer such question we can study the posterior predictive distributions. The comparison between the posterior predictive distribution with the observed data distribution gives us a visual idea about the quality of the fit. The comparison between the observed data histogram and the posterior predictive p.d.f of $X_{\max}$ given the composition inferred from data for the first energy bin is shown in FIGURE 5.2 .



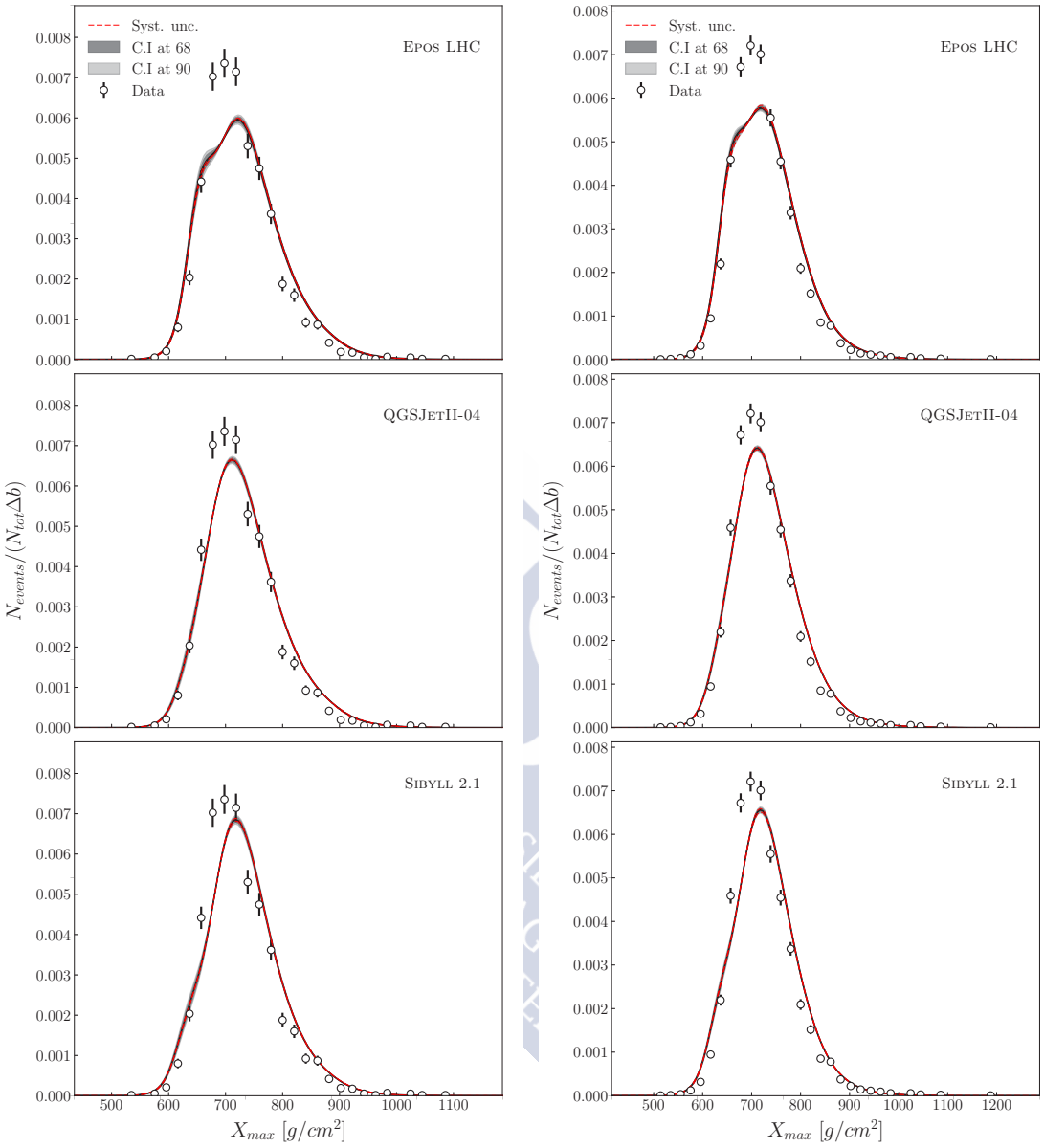
(A) With fiducial cuts.

(B) Without fiducial cuts.

FIGURE 5.2: Posterior predictive p.d.f for EPOS LHC (upper), QGSJETII-04 (middle) and SIBYLL 2.1 (lower) $X_{\max}$ in the energy bin $17.8 \leq \log_{10}(E/\text{eV}) < 17.9$.

Note that the posterior predictive p.d.f does not fit the observed $X_{\max}$ distribution in

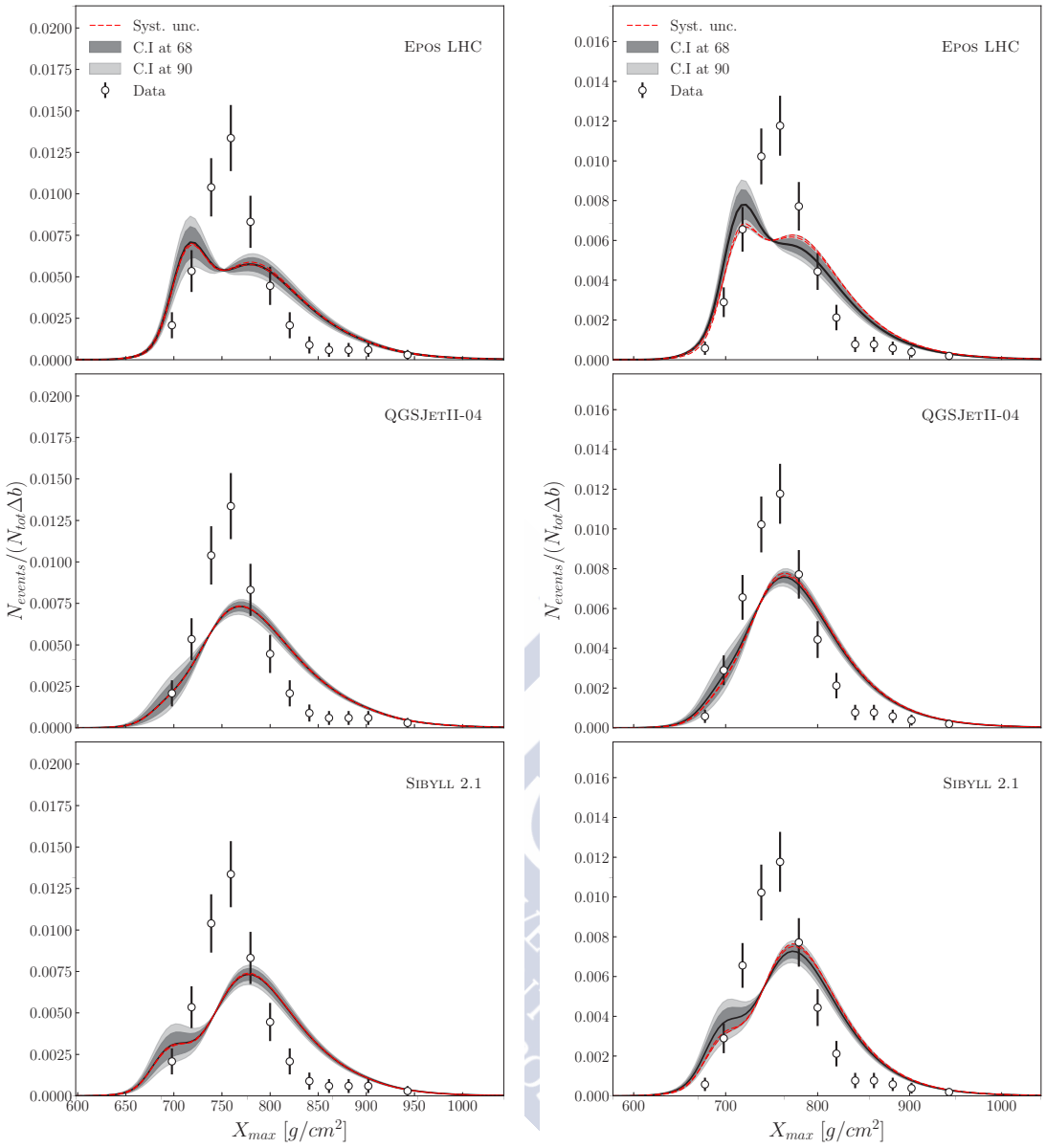
this range of energy. The same happens for the other energy bins (see FIGURES 5.3-5.4 for the comparison at log-energy bins $[18, 18.1]$ and $[19, 19.1]$).



(A) With fiducial cuts.

(B) Without fiducial cuts.

FIGURE 5.3: Posterior predictive p.d.f for EPOS LHC (upper), QGSJETII-04 (middle) and SIBYLL 2.1 (lower) $X_{\max}$ in the energy bin $18 \leq \log_{10}(E/\text{eV}) < 18.1$.



(A) With fiducial cuts. (B) Without fiducial cuts.

FIGURE 5.4: Posterior predictive p.d.f for EPOS LHC (upper), QGSJETII-04 (middle) and SIBYLL 2.1 (lower) $X_{\max}$ in the energy bin $19 \leq \log_{10}(E/\text{eV}) < 19.1$.

We can conclude that the observed $X_{\max}$ data by the Pierre Auger Observatory cannot be described by a p-Fe scenario with any of the three models considered. Therefore it becomes clear there is a need need to incorporate intermediate elements into the analysis. The obtained composition in the p-Fe scenario in this work is compared with the published in [66] (we are going to name [66] as AUGER12 from now on). Such comparison is shown in FIGURE 5.5. The agreement is clear within uncertainties.

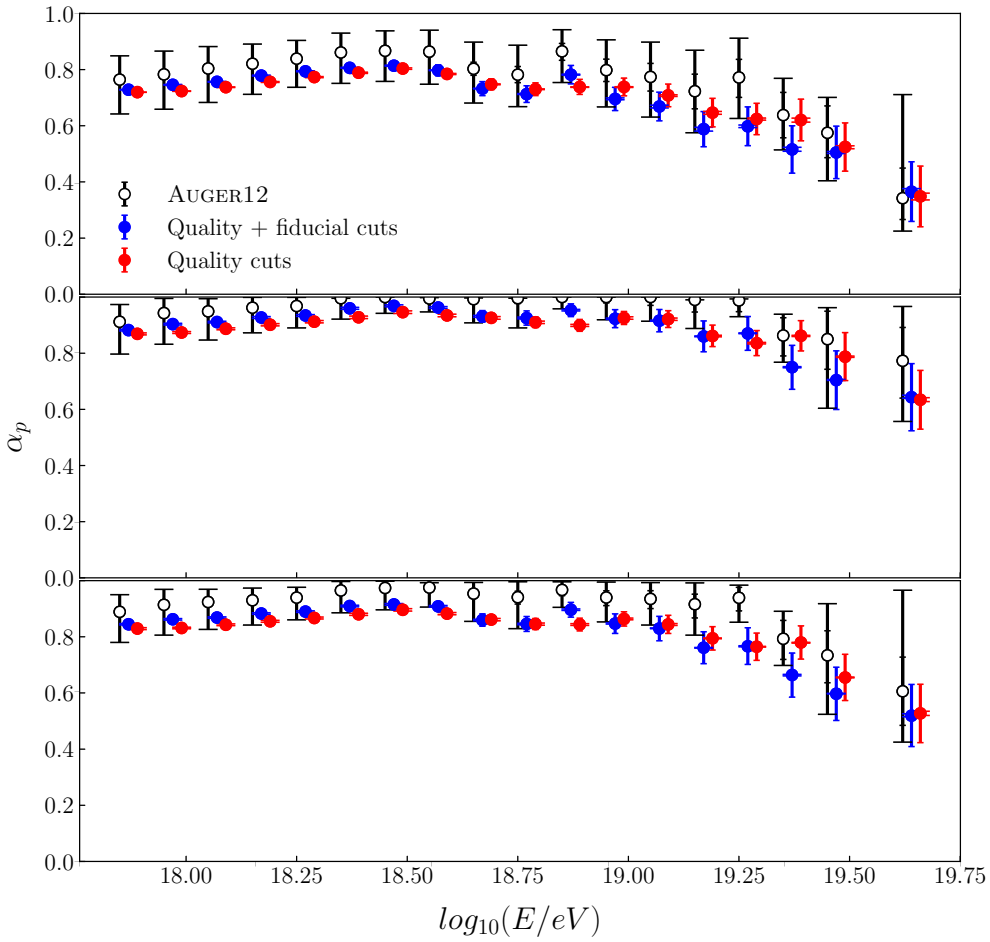


FIGURE 5.5: Comparison of the estimated fractions in AUGER12 (white circles) with those obtained in this work using anti-bias cut (blue circles) and without anti-bias cut (red circles) for the three hadronic models: EPOS LHC (upper panel), QGSJETII-04 (middle panel) and SIBYLL 2.1 (lower panel). The error bars with larger caps denote the systematic uncertainties, and those with smaller caps denote statistical uncertainties.

We can observe two main differences between the Bayesian analyses and the analysis performed by the Pierre Auger collaboration using a binned maximum likelihood (AUGER12) which is described in detail in [66]. On the one hand the composition inferred in the AUGER12 is systematically lighter. On the other hand while the statistical uncertainties are of the same magnitude the systematic uncertainties are very different.

The two approaches rely in a likelihood calculation but they are very different. In our case, the likelihood function is given by

$$\mathcal{L}(\alpha|D) = \prod_{i=1}^{N} \ell_i(X_{\max,i}), \tag{5.1}$$

where N is the total number of events and $\ell_i(X_{\max})$ is given by

$$\ell_i(X_{\max}) = \frac{\chi(X_{\text{low,i}}, X_{\text{up,i}}) \int_0^\infty \sum_j \alpha_j \mathcal{R}(X_{\max}|X^{true}_{\max}, E_i)\epsilon(X^{true}_{\max}|E_i) g_j(X^{true}_{\max}|E_i) dX^{true}_{\max}}{\int_0^\infty \chi(X_{\text{low,i}}, X_{\text{up,i}}) \int_0^\infty \sum_j \alpha_j \mathcal{R}(X_{\max}|X^{true}_{\max}, E_i)\epsilon(X^{true}_{\max}|E_i) g_j(X^{true}_{\max}|E_i) dX^{true}_{\max} dX_{\max}} \tag{5.2}$$

All the terms that appear in EQUATION 5.2 have been explained in previous chapters (see SECTION 3.7). Then, applying the Bayes' theorem we obtain the posterior p.d.f of the fractions.

In the AUGER12, for each energy bin the observed $X_{\max}$ distribution is binned in $X_{\max}$ ranges of 20 g/cm^2. The likelihood function is given by

$$\mathcal{L} = \prod_{b=1}^{B} \ell_b, \tag{5.3}$$

where B is the number of bins of the $X_{\max}$ range and ℓ_b is given by

$$\ell_b = \frac{e^{-C_b} C_b^{n_b}}{n_b!}. \tag{5.4}$$

This is the probability of obtaining n_b events when C_b are expected assuming n_b is Poissonian. The expected number of events that falls in bin b that belongs to primary j given the the fraction α_j is:

$$C_b = N_{data} \sum_j \alpha_j G_{j,b} \tag{5.5}$$

Here, $G_{j,b}$ is the fraction of events that fall in bin n and it is calculated from simulations for each primary. Notice that this approach is good enough if the number of events is large but it is somewhat limited if there is a small number of events in the data sample. In addition it introduces additional systematic uncertainties due to the chosen binning.

5.2 Other 2-primaries scenarios

We have concluded that p and Fe are not sufficient to describe the data. It is now possible to address the question of whether protons or irons are really needed. To answer this question more scenarios with two primaries have been studied: p-He, p-N, He-N, He-Fe and N-Fe. The trends of the composition fractions for all these hypotheses are shown in Appendix C. For all the scenarios Epos LHC continues being the hadronic model which gives the heaviest composition and QGSJetII-04 gives the lightest composition. As in the previous section, there are no significant differences between analysing using the data sample passing both quality and fiducial cuts or using the data sample passing only the quality cuts. As an example, the proton evolution in the p-N scenario for is shown in Figure 5.7

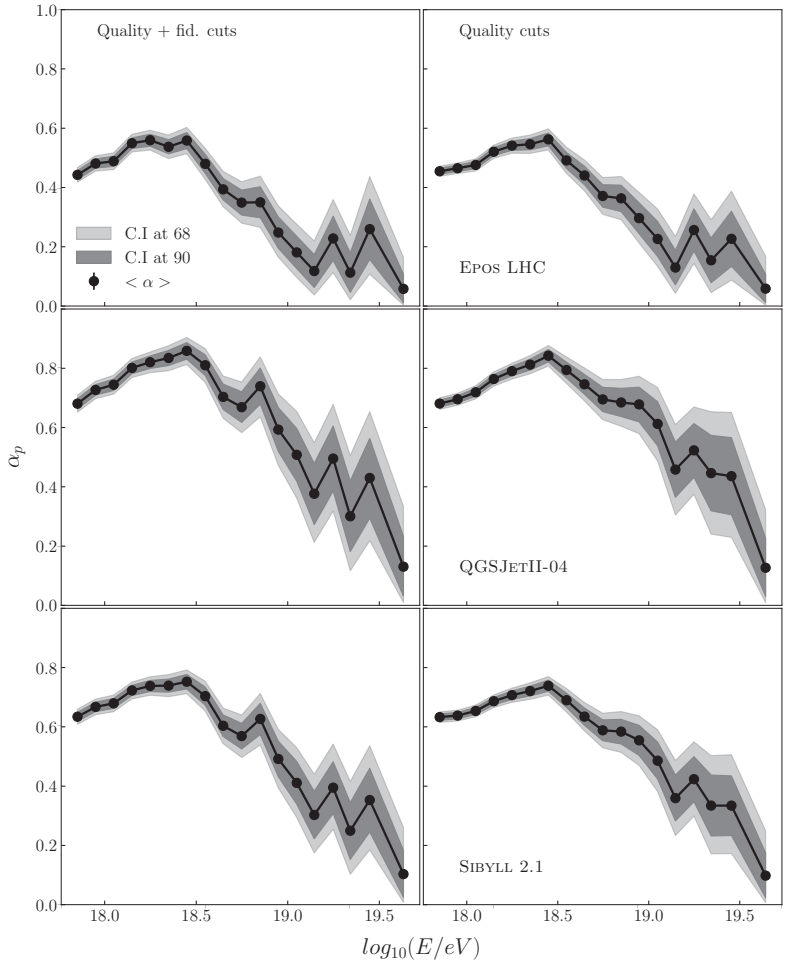


Figure 5.6

Figure 5.7: Proton fraction as a function of the energy using Sibyll 2.1 hadronic interaction model in the p-N scenario applying quality + fiducial cuts (A) and applying only quality cuts (B).

It is clear that if we analyse the data assuming different scenarios we will obtain different composition fractions. In FIGURE 5.8 we display the mean values of the posterior p.d.f of the fraction of the lightest element element using EPOS LHC model for the 6 scenarios with only two primaries. In almost all scenarios one can see a pattern in the behaviour of the fraction of the lightest element, particularly in the region where there are more events and we have less uncertainty.

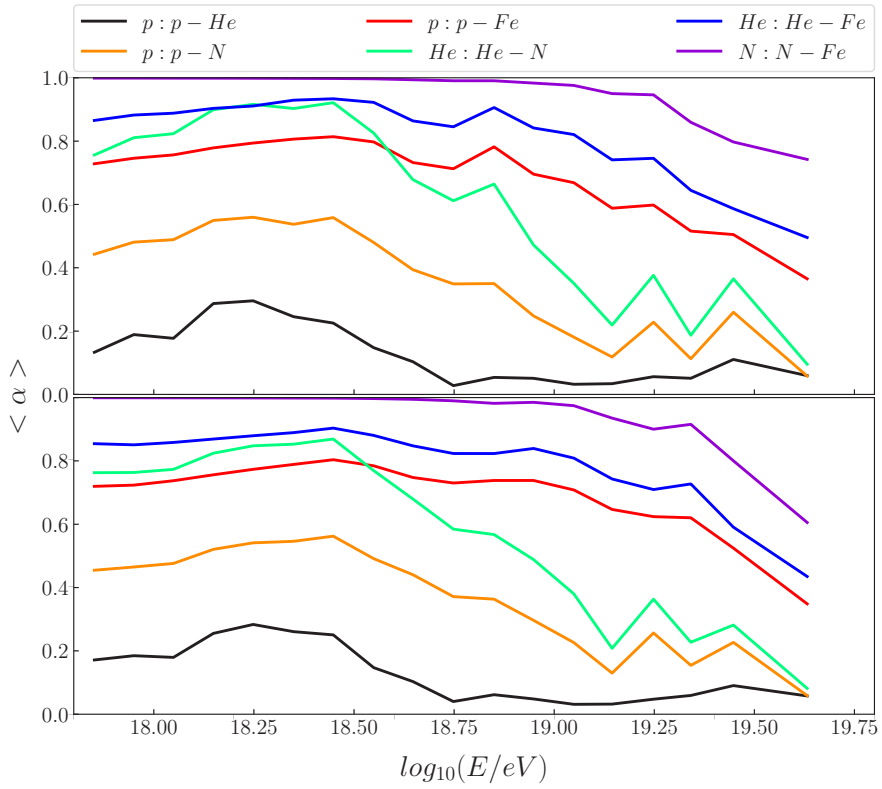


FIGURE 5.8: Mean values of the posterior p.d.f of the lightest primary for the different scenarios using EPOS LHC hadronic interaction model: p-He (black), p-N (orange), p-Fe (red), He-N (green), He-Fe (blue) and N-Fe (violet). In the upper panel the inference is done using the data sample with both quality and fiducial cuts applied. In the lower panel only the quality cuts are applied in the data sample.

Note that comparing results using scenarios with only two primaries some important conclusions can be obtained. If we assume only p and He any primary which is heavier would be assigned to He. As we obtain a significant fraction of protons in this scenario we can conclude that there are protons (assuming that the hadronic model is correct[3]). As we change the composition hypothesis to p-N or p-Fe we obtain larger proton fractions indicating that intermediate masses between protons

[3]Here we of course neglect the possibility of any isotope with $A = 2$ or $A = 3$.

and nitrogens or irons are needed. If we consider the N-Fe scenario we obtain nearly 100% of nitrogen being the fraction of Fe close to zero below 10 EeV. In the He-N scenario the fraction of He is high but not 100%. The results of these two scenarios (N-Fe and He-Fe) mean that the presence of irons could be negligible with respect to the other primaries. Finally, by analysing He-N scenario we see that the helium fraction rises a little up to $\log_{10}(E/\text{eV}) = 18.5$ and drops for larger energies from 90% to 20% at $\log_{10}(E/\text{eV}) = 19.2$. The drop is stronger for He-N than for He-Fe. This implies that the composition is getting heavier and that intermediate mass nuclei such as N is needed.

Using this reasoning and taking into account that Epos LHC hadronic model is which gives the heaviest composition, we can conclude that protons are needed at least up to 10 EeV and that intermediate masses are also needed. This results agree with those discussed in [67].

Nevertheless it is through the posterior odds that we can analyse quantitatively the different scenarios (see Section 2.6.4) or, if we have more than two scenarios, through Equation 2.46 which we rewrite here:

$$P(H_m|D) = \frac{P(D|H_m)P(H_m)}{P(D)} = \frac{Z_m P(H_m)}{\sum_{l=1}^{M} Z_l P(H_l)}. \tag{5.6}$$

In Equation 5.6, H_m is a scenario, $Z_m = P(D|H_m)$ is its evidence and $P(H_m)$ is its prior probability. In our analysis the prior probabilities for all scenarios are equal. The number of hypothesis M is 18 which corresponds with 6 different assumptions for 2 primary mixtures times three different models: $\{p-He, p-N, p-Fe, He-N, He-Fe, N-Fe\} \times \{$Epos LHC ,QGSJetII-04 ,Sibyll 2.1 $\}$. Now we proceed to compare the different cases.

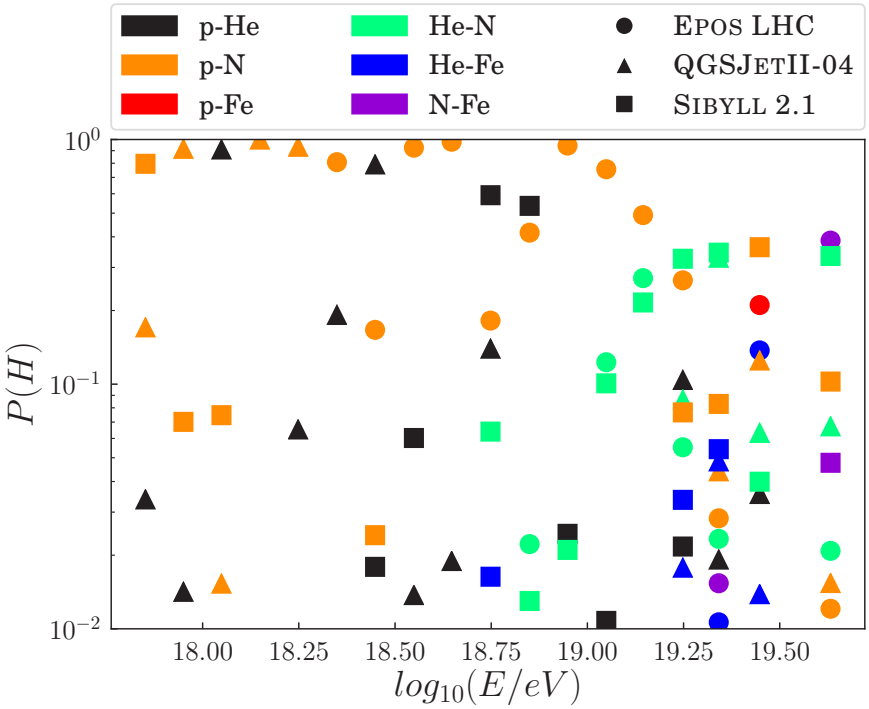


FIGURE 5.9: Probability of the 18 different hypotheses in the bi-component case. The different composition scenarios are differentiated by colours: black (p-He), orange (p-N), red (p-Fe), green (He-N), blue (He-Fe) and violet (N-Fe). The different hadronic interaction models are differentiated using different markers: circles (EPOS LHC), triangles (QGSJETII-04) and squares (SIBYLL 2.1). Only hypothesis with a probability greater than 10^{-2} are shown for better viewing.

In the above figure all hypotheses are shown. Clearly there is no preferred hypothesis to describe all the data. One can observe that up to 10 EeV the preferred hadronic interaction model depends of the bin but always there are protons in the preferred model, sometimes mixed with nitrogen and other times mixed with helium. Beyond 10 EeV we can see a change to heavier elements but the differences among the probabilities of the different scenarios also decreases since the number of events decreases with the energy. At these energies the data have little power to give relevant conclusions.

To get a visual idea about how the probabilities of the hypotheses are translated to the data, in FIGURE 5.10 the posterior predictive $X_{\max}$ distributions for the energy bin $18 \leq \log(E/\text{eV}) < 18.1$ in, for instance, the p-He scenario. In this scenario the probabilities of these hypotheses are approximately 10^{-64}, 0.91 and 10^{-12} for EPOS LHC, QGSJETII-04 and SIBYLL 2.1 respectively.

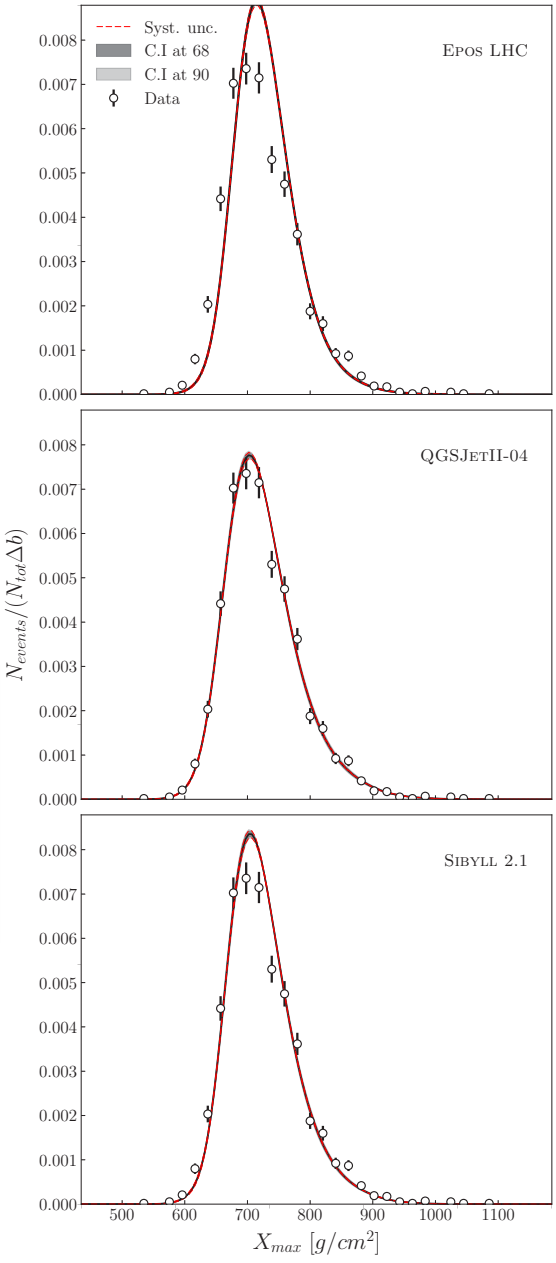


FIGURE 5.10: $X_{\max}$ posterior predictive p.d.f for the energy bin $18 \leq \log(E/\text{eV}) < 18.1$ in the using p-He to fit the composition. The probabilities of these hypotheses are $6 \cdot 10^{-64}$, 0.91 for EPOS LHC (upper panel), 0.91 for QGSJETII-04 and 10^{-12} for SIBYLL 2.1.

Note that while QGSJETII-04 seems to fit quite well the $X_{\max}$ distribution using only proton and helium in this energy bin, EPOS LHC model cannot fit well the left tail and needs more heavier primaries. SIBYLL 2.1 fits better the $X_{\max}$ distribution than EPOS LHC but worse than QGSJETII-04.

Now we are going to study the hypotheses separating the hadronic interaction models to infer a possible composition trend as a function of energy. Remember that we are studying a possible trend for each of the four primaries but analysing the data samples assuming only two primaries.

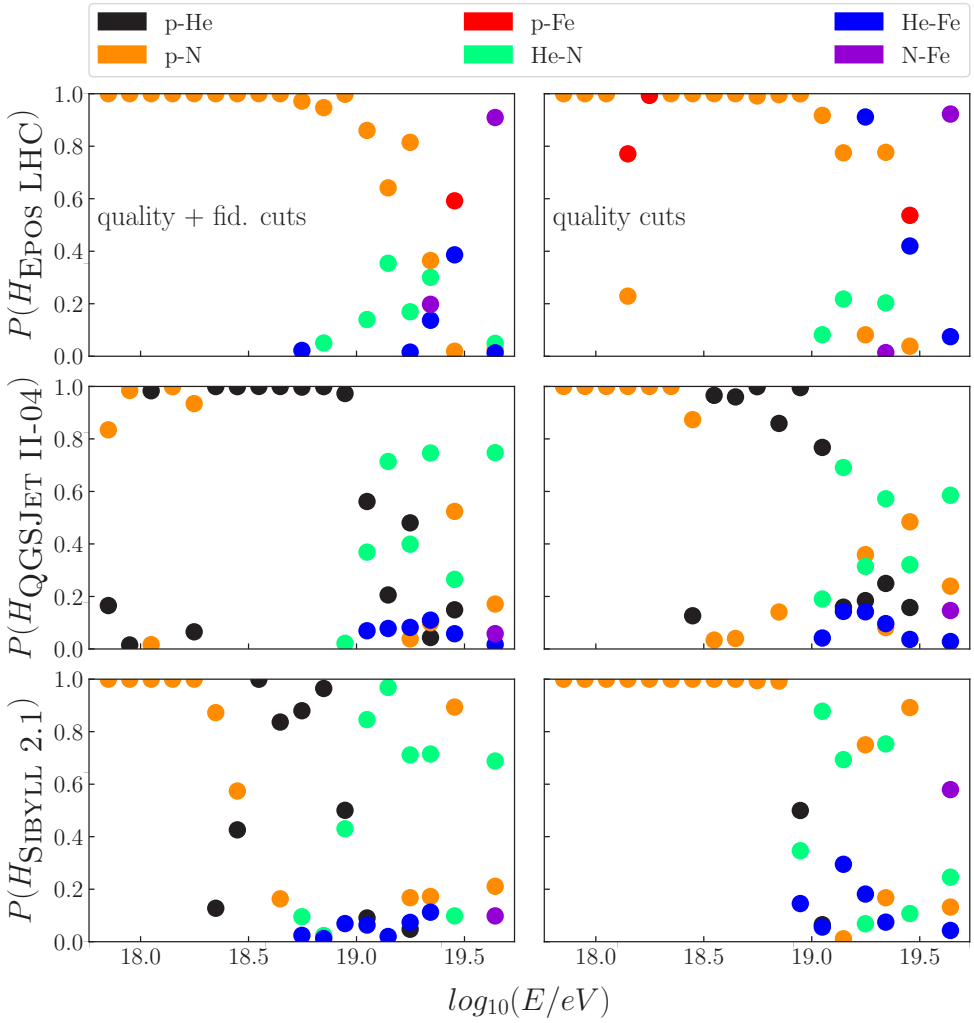


FIGURE 5.11: Probability of the various hypotheses as a function of energy organised by hadronic interaction model and data sample. In upper panels we consider EPOS LHC while QGSJETII-04 and SIBYLL 2.1 are respectively shown in the middle and lower panels. At right the anti-bias cut has been applied over the data. In right panels only the quality cuts have been applied. to the data and in the right panels the antibias cut has also been applied. The colours for distinguishing the different composition scenarios are the same as those adopted in FIGURE 5.8.

Only hypotheses with a probability greater than 10^{-2} are shown for clarity.

By looking at FIGURE 5.11 and focusing on the analysis performed over the data sample with the anti-bias cut applied using EPOS LHC model we can observe that up

to 10 EeV the most probable scenario is p-N. At energies around 20 EeV it seems to be a change in the behaviour and the models containing He are not irrelevant. Beyond these energies the scenarios with presence of iron begin to have some importance. The fractions using QGSJetII-04 and Sibyll 2.1 are similar but differ from those obtained with Epos LHC. Up to energies around 3.2 EeV the most probable scenario is p-N. In the range $[3.2 - 10]$ EeV the scenario changes into p-He and beyond these energies the He-N scenario seems to be the most probable one.

Notice that by using only combinations of two-component scenarios we are observing trends in changes of composition with energy for four primaries through the posterior odds. Nevertheless, if we want to extract more information about these trends we must analyse the data using more primaries and not just two.

We are going to compare the analyses of the data samples with and without the anti-bias cut. As was discussed in Section 3.7.2 both analyses should result in the same inferences as long as the efficiency and response functions of our detector are well described. We are going to check this comparing the outcome from the two analyses. We can see that the probabilities of the two sets of data are in reasonable agreement but they are not exactly the same (left and right panels of Figure 5.11. For the three hadronic models we can observe that in some energy bins the most probable scenarios become heavier when we do not apply fiducial cuts. For example, the most probable scenario using Epos LHC hadronic model in the energy range $18.2 \leq \log_{10}(E/\text{eV}) < 18.4$ is p-Fe while when we use the data with anti-bias cut applied the most probable model is p-N. Using Sibyll 2.1 the p-He scenario practically disappears when the anti-bias cut is applied and a similar behaviour occurs when we use the QGSJetII-04 model but the change is not so drastic. These differences could be due to different reasons: our composition scenarios are too simple or we need to add more primaries in each energy bin, the hadronic models do not reproduce the actual high energy interactions or the detector is not well characterised. To investigate the former, in the next section we analyse the data samples using three primaries.

To finish this section let us go back to the question: are protons and irons needed? In sight of Figure 5.11 it seems clear that the presence of protons is needed for all the hadronic interaction models, particularly at energies up to 10 EeV. For higher energies we cannot discard the presence of proton but it seems that its presence becomes less important. The presence of iron seems to be necessary mainly when analysing with Epos LHC but this only happens at the highest energies, where we have less number of events and our inference is thus subject to more uncertainties.

5.3 Scenarios with 3 primaries

To analyse data with three primaries we are going to assume four composition scenarios: p-He-N, p-He-Fe, p-N-Fe and He-N-Fe. A total of 12 hypotheses are going to be analysed (4 composition scenarios multiplied by 3 hadronic models). Note that all the hypotheses assumed in the previous section are subsets of those of this section.

In FIGURE 5.11 the probability of the all scenarios is shown. For instance, while SIBYLL 2.1 is used, the most probable scenario is p-N at lower energies but there was a transition to p-He and then another transition to heavier elements as the energy increases. For this reason and in order to understand this result we are going to show the results for the p-He-N scenario using SIBYLL 2.1. The trends of the composition fractions are shown in FIGURE 5.12.

In this scenario the most predominant primary at lower energies is proton followed by nitrogen. From 1 EeV up to $\log_{10}(E/\text{eV}) = 18.5$ the proton fraction initially increases and then drops (like in the previous section) and it continues decreasing up to 10 EeV. Once more a local maximum is found in the energy range $\log_{10}(E/\text{eV}) \in [18.2, 18.4]$. Above this energy, as the proton fraction decreases the helium fraction rapidly grows up to energies around 10 EeV when it reaches a maximum and it starts to fall. At higher energies, as the helium fraction is dropping the nitrogen fraction increases to become the most abundant element at the highest energies.

Note that this trend in energy of the composition is totally in agreement which the probability of the 2-primary scenarios shown in FIGURE 5.11 and discussed in SECTION 5.2, where it was shown that those that better describes the data was p-N at lower energies, p-He at intermediate energies and He-N at higher energies.

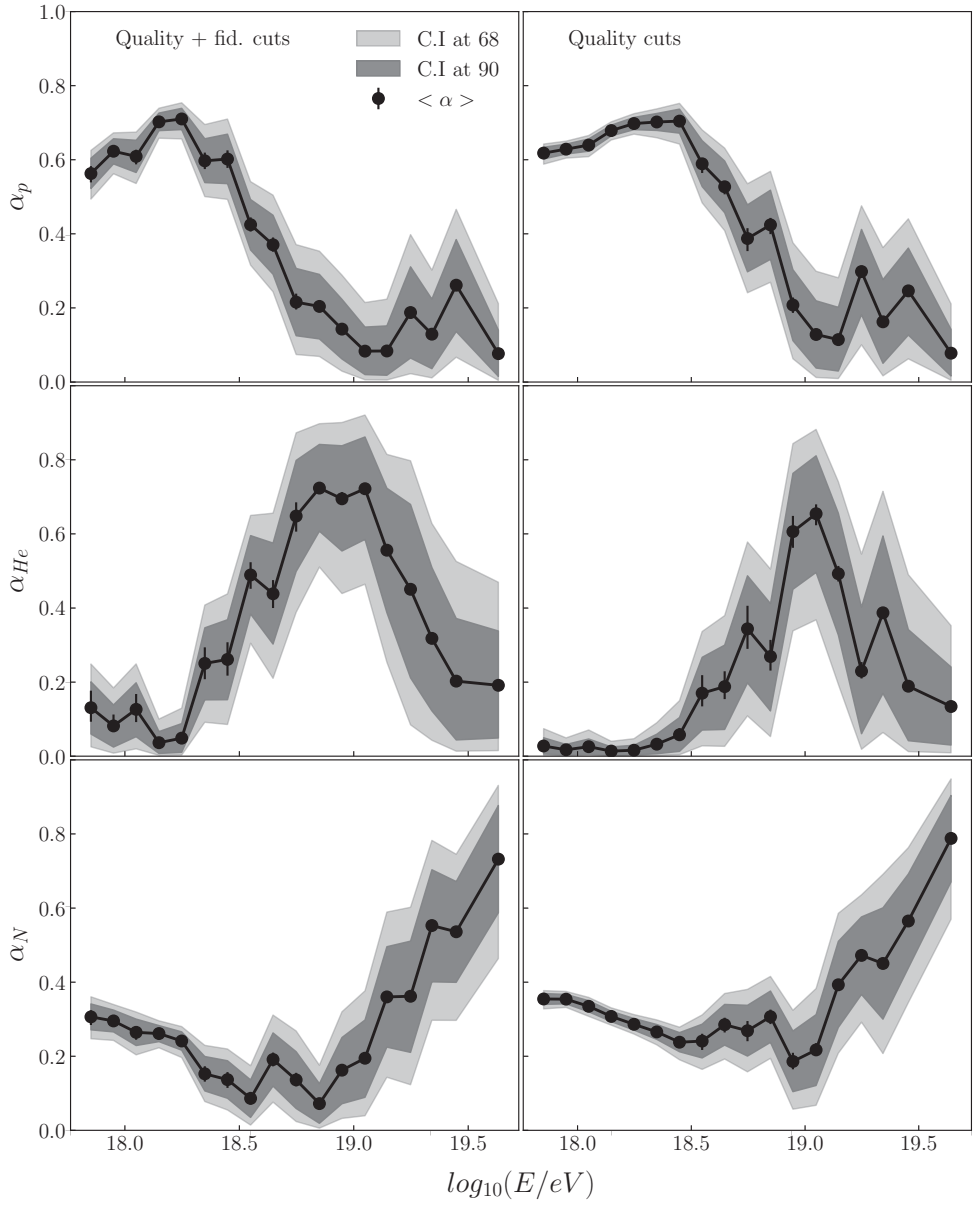


FIGURE 5.12: pr-He-N evolution using SIBYLL 2.1 with the data sample with anti-bias cut (left) and without anti-bias cut (right). Systematic uncertainties are represented with error bars. The 68% and 90% of confidence interval of the posterior are shown as shaded bands.

If we take a look at the analysis using the data sample without the anti-bias cut we observed a similar conclusions that in the previous section. The uncertainties using the data set without the anti-bias cut are smaller than when the anti-bias cut is used. Up to $\log_{10}(E/\text{eV}) = 18.5$ the helium fraction is practically null but beyond this energy it increases reaching a maximum at an energy around 10 EeV. Above this energy the composition becomes heavier.

The differences between the composition inferred using anti-bias cuts and without anti-bias cuts could be due to the mentioned reasons in the previous section but could also be due to the different number of events in the data samples. Note that when the anti-bias cut is not applied the number of events increases and a better inference is expected.

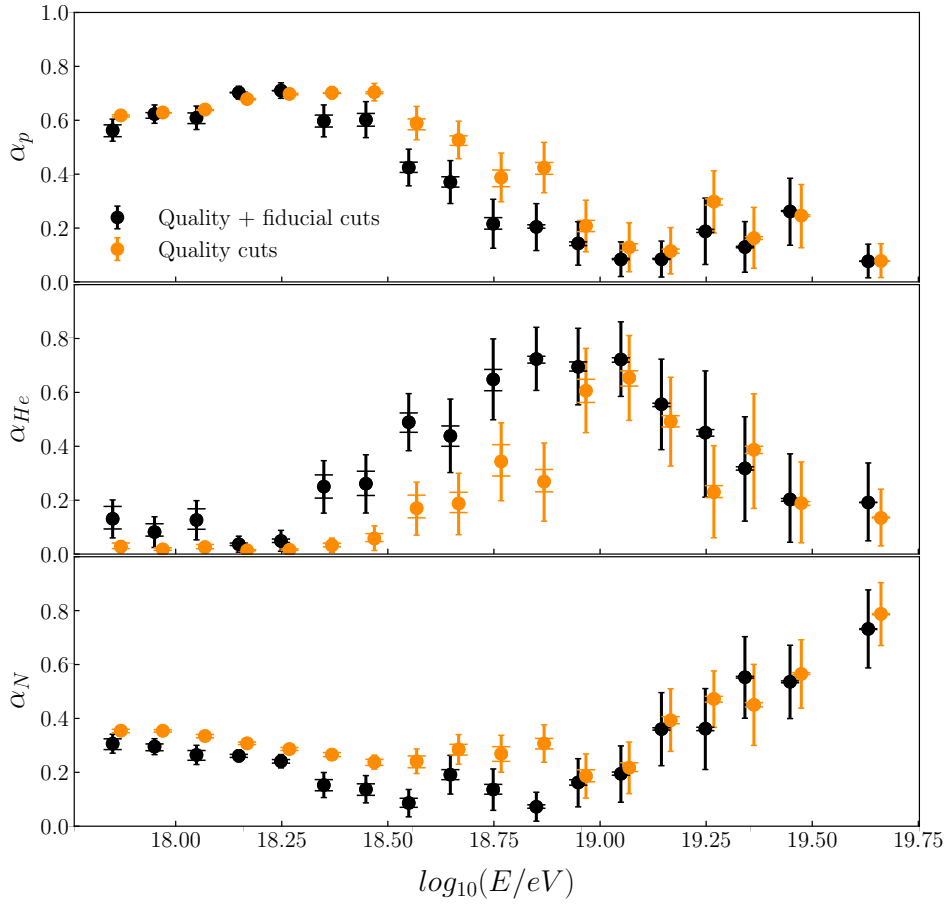


FIGURE 5.13: Inferred composition using SIBYLL 2.1 hadronic interaction model in the p-He-N scenario. The black points correspond with the mean value of the posterior p.d.f using the anti-bias cut. Orange is for the inferences using data without anti-bias cut. The error bars with smallest caps denote the 68% of confidence interval and the highest caps for systematic uncertainties. The proton evolution is shown in the upper panel, helium and nitrogen trends are shown in the middle and lower panels respectively. Energies for data without anti-bias cut have been shifted for better viewing.

The largest deviations between the analyses using fiducial cuts and without fiducial cuts occurs when the systematic uncertainties are also the largest. One can see that

the helium fraction is most affected by the systematic uncertainties. This effect should not come as surprise because it is already known that the helium $X_{\max}$ distribution is the primary with the smallest non-overlapped area (see SECTION 3.5). Then, its inference subject to more uncertainty and helium events can be misidentified sometimes as protons and other times as nitrogens or viceversa. The composition trends inferred using EPOS LHC and QGSJETII-04 are shown in FIGURES 5.14-5.15.

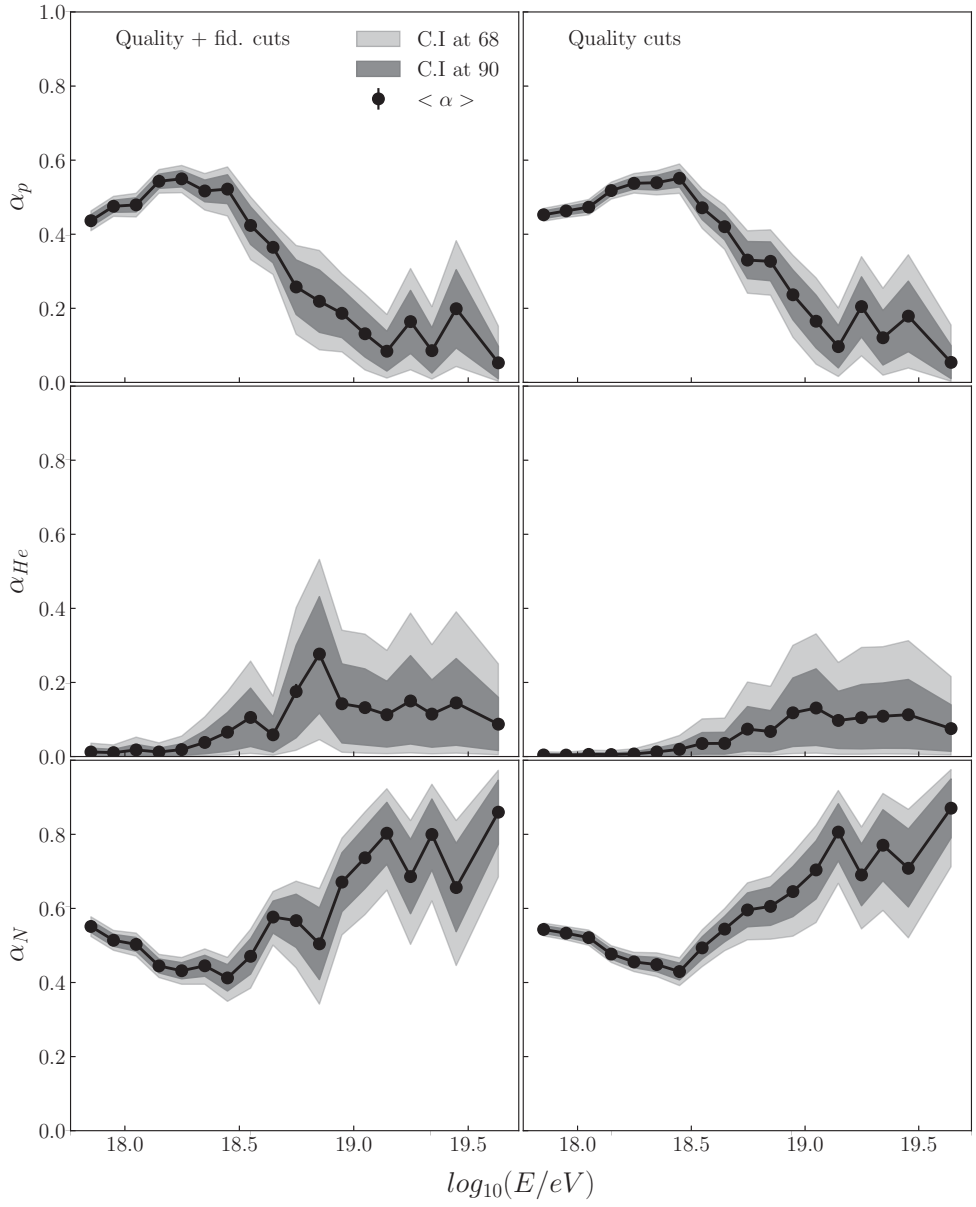


FIGURE 5.14: Same as FIGURE 5.12 but using EPOS LHC hadronic interaction model.

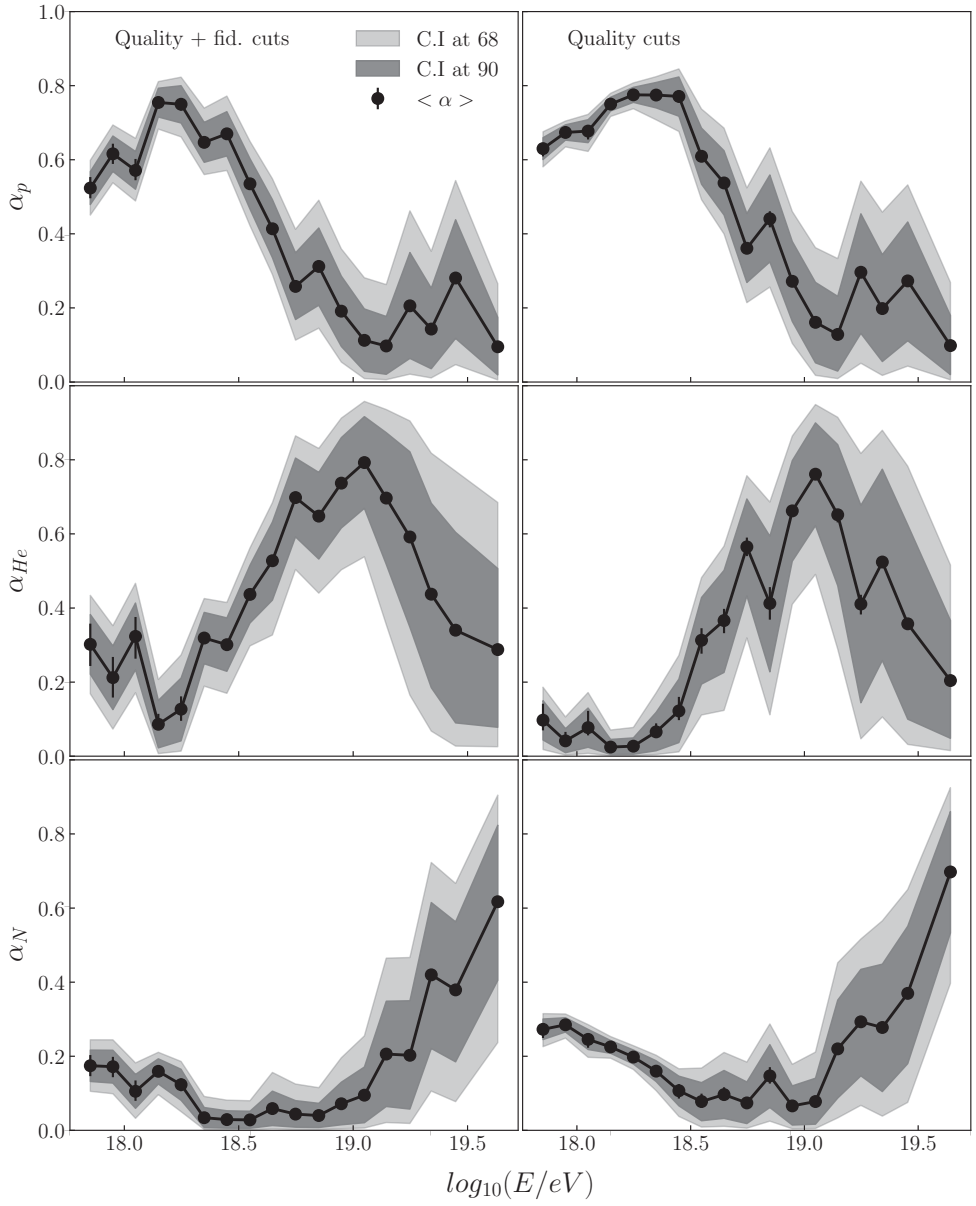


FIGURE 5.15: Same as FIGURE 5.12 but using QGSJETII-04 hadronic interaction model.

In the three hadronic interaction models we observe at lower energies the same transition that we discussed in the previous section for almost all two-component scenarios. The composition becomes lighter up to $\log_{10}(E/\text{eV}) = 18.4$ when proton fraction begins to drop as heavier elements take in. There are differences in the absolute values of the fractions for the different models particularly between EPOS LHC and the other hadronic models.

While both SIBYLL 2.1 and QGSJETII-04 show a clear transition from protons to heliums beyond the local maximum of the proton fraction, using EPOS LHC this

transition is from protons to helium and nitrogen (or even to nitrogen only). The data set needs a large presence of nitrogen if it is analysed with EPOS LHC.

We now analyse the p-N-Fe scenario. The composition inferred here can be compared with that of AUGER12 [66]. Such comparison is shown in FIGURES 5.16-5.18.

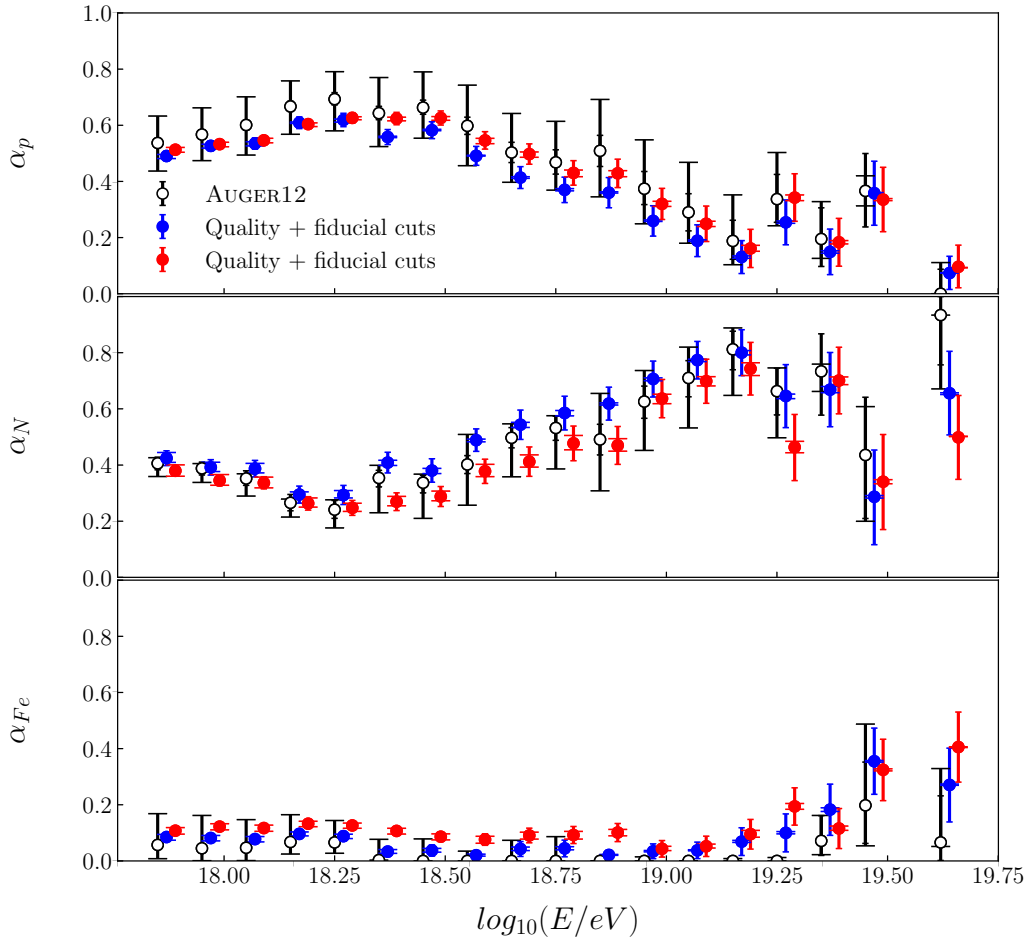


FIGURE 5.16: Comparison of the estimated fractions in AUGER12 [66] (white circles) with those obtained in this work using anti-bias cut (blue squares) and without anti-bias cut (red squares) using the EPOS LHC hadronic interaction model. In the upper, middle and lower panels the proton, nitrogen and iron compositions are shown respectively. The error bars represent the 68% of confidence level.

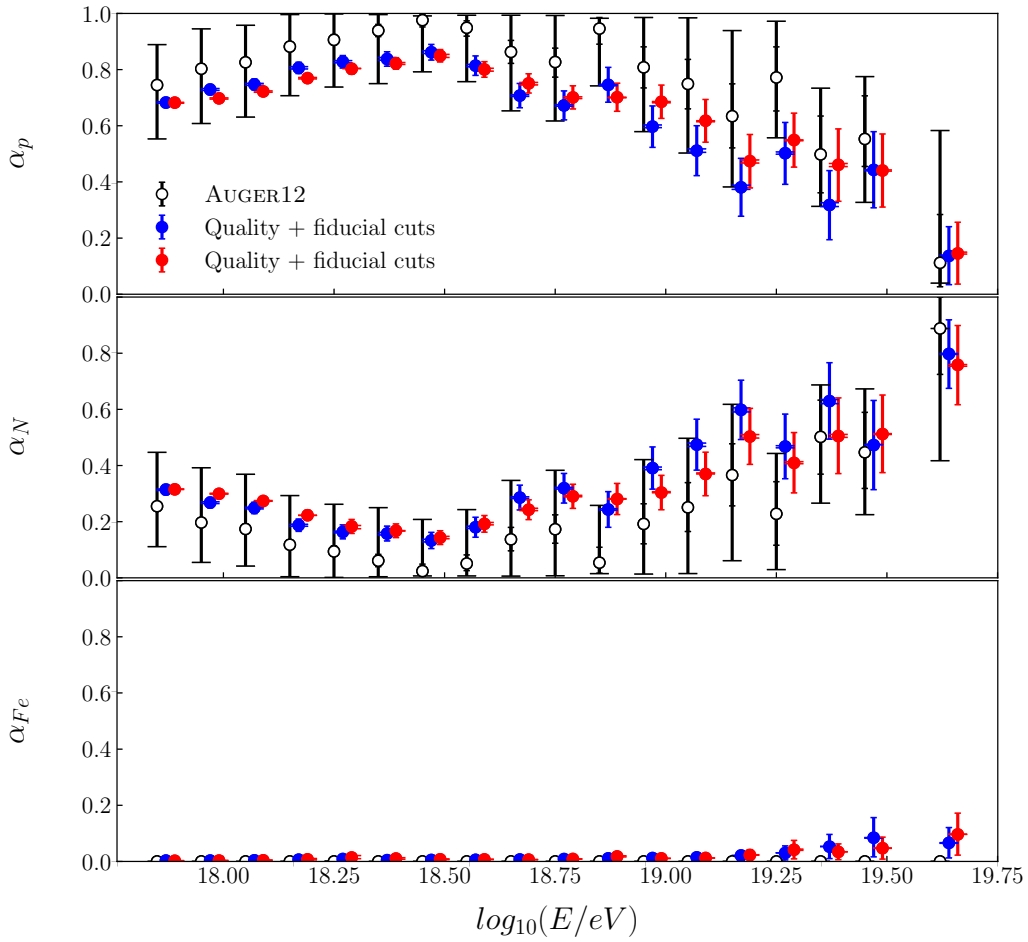


FIGURE 5.17: Same as FIGURE 5.16 but using QGSJETII-04 model.

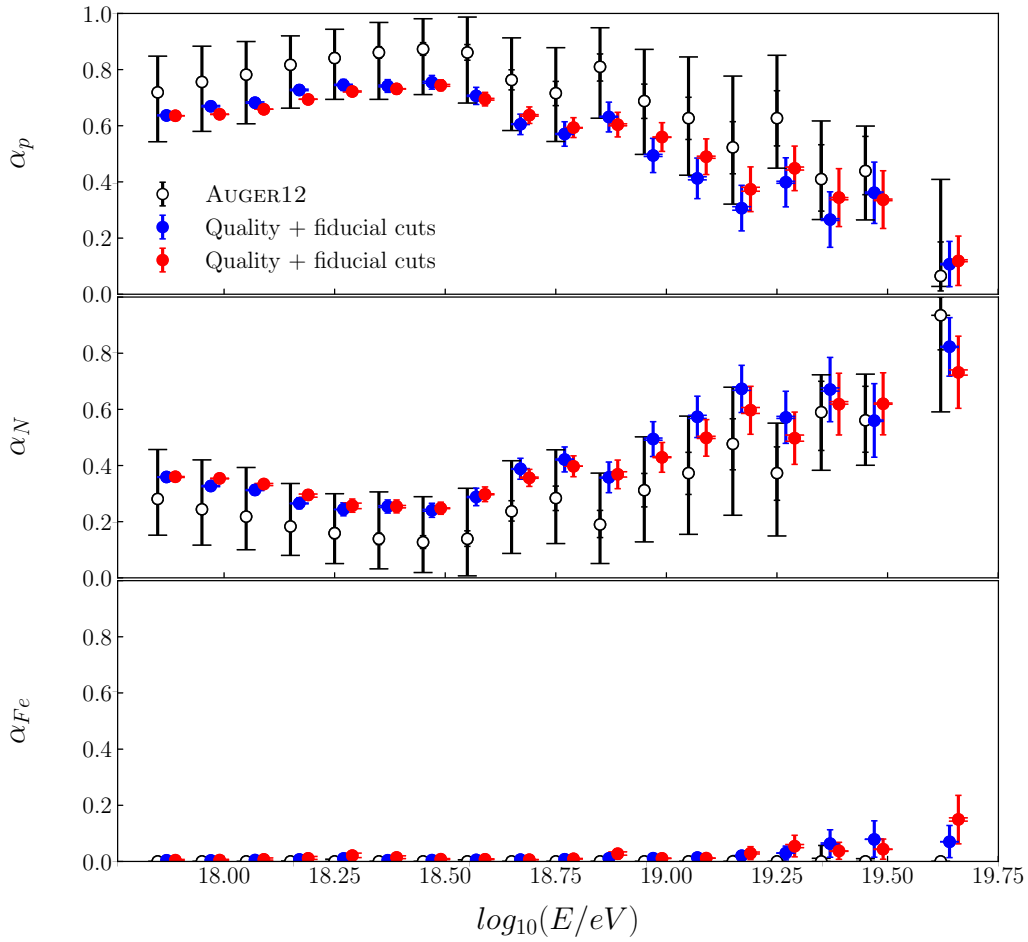


FIGURE 5.18: Same as FIGURE 5.16 but using SIBYLL 2.1 model.

As mentioned above the EPOS LHC model gives us heavier composition than any of the other models. In fact, using EPOS LHC the iron fraction is not null at lower energies while with the other models iron shows up only at the highest energies. The composition fractions inferred in AUGER12 [66] and those inferred in this work differ more in this scenario than in the p-Fe scenario. Although the fractions obtained in both analyses are compatible. We note that the AUGER12 composition is lighter than the composition inferred using the Bayesian methods in both data samples (with and without anti-bias cut).

For SIBYLL 2.1 and QGSJETII-04 there are no significant differences between the analyses performed using the anti-bias cut or without it while for EPOS LHC these differences are larger. Once more we can see that the composition becomes lighter as

the energy increases reaching a maximum (which depends on the hadronic interaction model) and then it becomes heavier.

As in the previous section we can calculate the probability of each scenario. These probabilities are shown if FIGURE 5.19. As happened in the two-component scenarios, at lower energies if EPOS LHC is the most probable scenario the inference gives heavier composition than when QGSJETII-04 or SIBYLL 2.1 are the preferred scenarios. The He-N-Fe scenario is very disfavoured in all the energy range except for the largest energies.

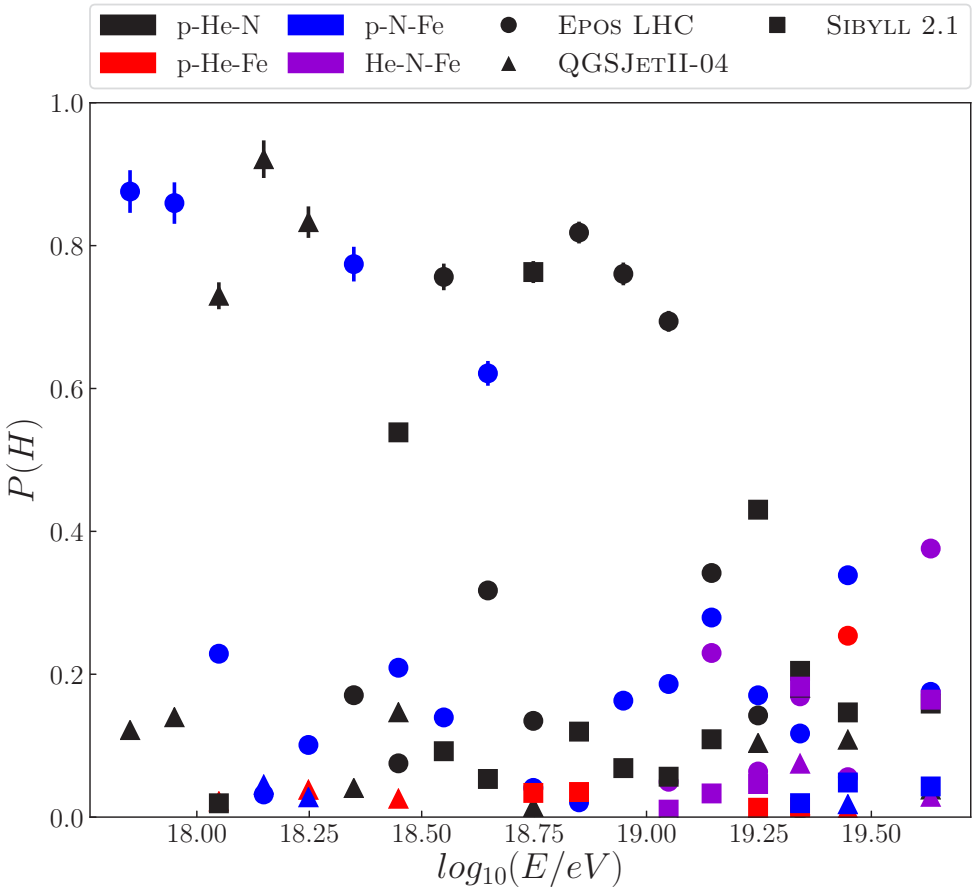


FIGURE 5.19: Probability of the 12 different cases in the three-component scenarios.Black (p-He-N), red (p-He-Fe), blue (p-N-Fe) and violet (He-N-Fe), circles (EPOS LHC), triangles (QGSJETII-04) and squares (SIBYLL 2.1). Only hypotheses with a probability greater than 10^{-2} are shown for better viewing.

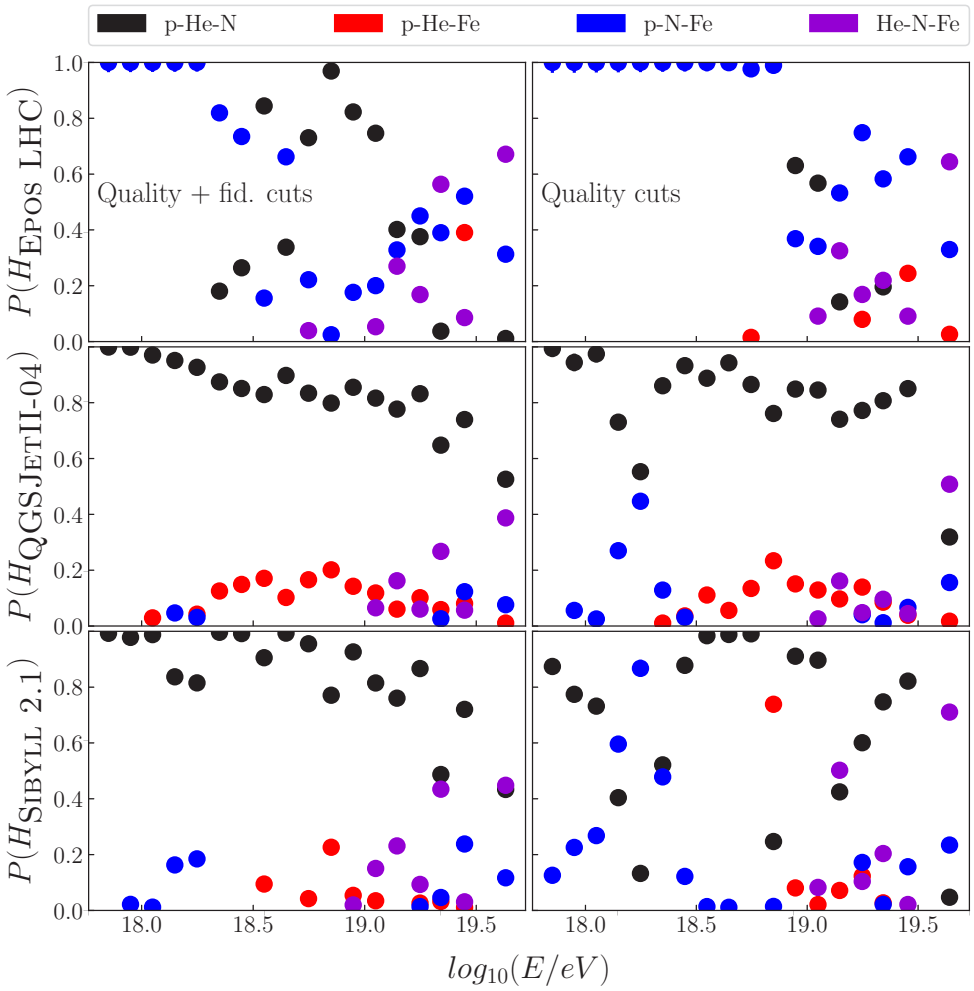


FIGURE 5.20: Probability of the different three-primary scenarios for EPOS LHC (upper raw), QGSJETII-04 (middle raw) and SIBYLL 2.1 (lower raw) for data with anti-bias cut (left panels) and without anti-bias cut (right panels). The primary scenarios are differentiated with the same colours as FIGURE 5.19. As in FIGURE 5.11, scenarios with a probability less than 10^{-2} are not shown.

The probabilities of the four three-component scenarios for the three hadronic interaction models using data with and without anti-bias cut are shown in FIGURE 5.20. The conclusions are similar to those of the two-component scenarios in FIGURE 5.11. When the fiducial cut is removed heavier compositions are favoured. This effect is more apparent when the analysis is performed using EPOS LHC. Although this effect can also be seen when analysing with QGSJETII-04 , this model seems to be the least affected by it. The p-He-N keeps on being the most probable scenario in all the energy range except in the last energy bin when the most probable scenario becomes in He-N-Fe.

5.4 p-He-N-Fe scenario

Up to now we have been working with protons, heliums, nitrogens and irons in separated ways. First, combining pairs of primaries and later combining tree primaries. In this section we analyse the data samples assuming that the four primaries. In this analysis we use protons as the lightest element, the helium distribution represents the low masses, the nitrogens represent intermediate masses and the iron distribution represent highest masses. The comparisons between the analyses performed using the anti-bias cut and without the anti-bias cut for the different hadronic models are shown in FIGURES 5.21-5.23.

The addition of a fourth element to the analysis does not affect qualitatively the energy trend of the proton fraction. For the three models we observe at the lower energies an increase of the proton fraction. Protons reach a maximum at $\log_{10}(E/\text{eV}) \approx 18.3$ and then fall. This behaviour does not depend on the hadronic model used and it is the same when we analyse data with or without fiducial cuts. It is interesting to remark once more that the absolute values of the composition fractions do not change much when we compare the analyses of data with and without anti-bias cut (as in the previous sections).

The comparison of the estimated fractions obtained using different hadronic models is more interesting. When SIBYLL 2.1 or QGSJETII-04 is used in the analysis the helium fraction decreases as the proton fraction increases (the nitrogen fraction also decreases). When the proton fraction reaches its maximum, the helium fraction reaches a local minimum and there is a transition from protons to heliums. At energies around $18.8 \leq \log_{10}(E/\text{eV}) < 19$ there is a local maximum in the helium fraction and above this energy the fraction falls. As a result of the decrease of the helium fraction the nitrogen fraction increases (and possibly the iron fraction too). This sequence of the decreasing fractions of lighter elements in favour of heavier elements as the energy rises can be interpreted as an effect due to Peters' cycle (see [83]). In this scenario the maximum energy that a source can achieve is proportional to the nucleus charge, hence heavier nuclei would reach higher energies.

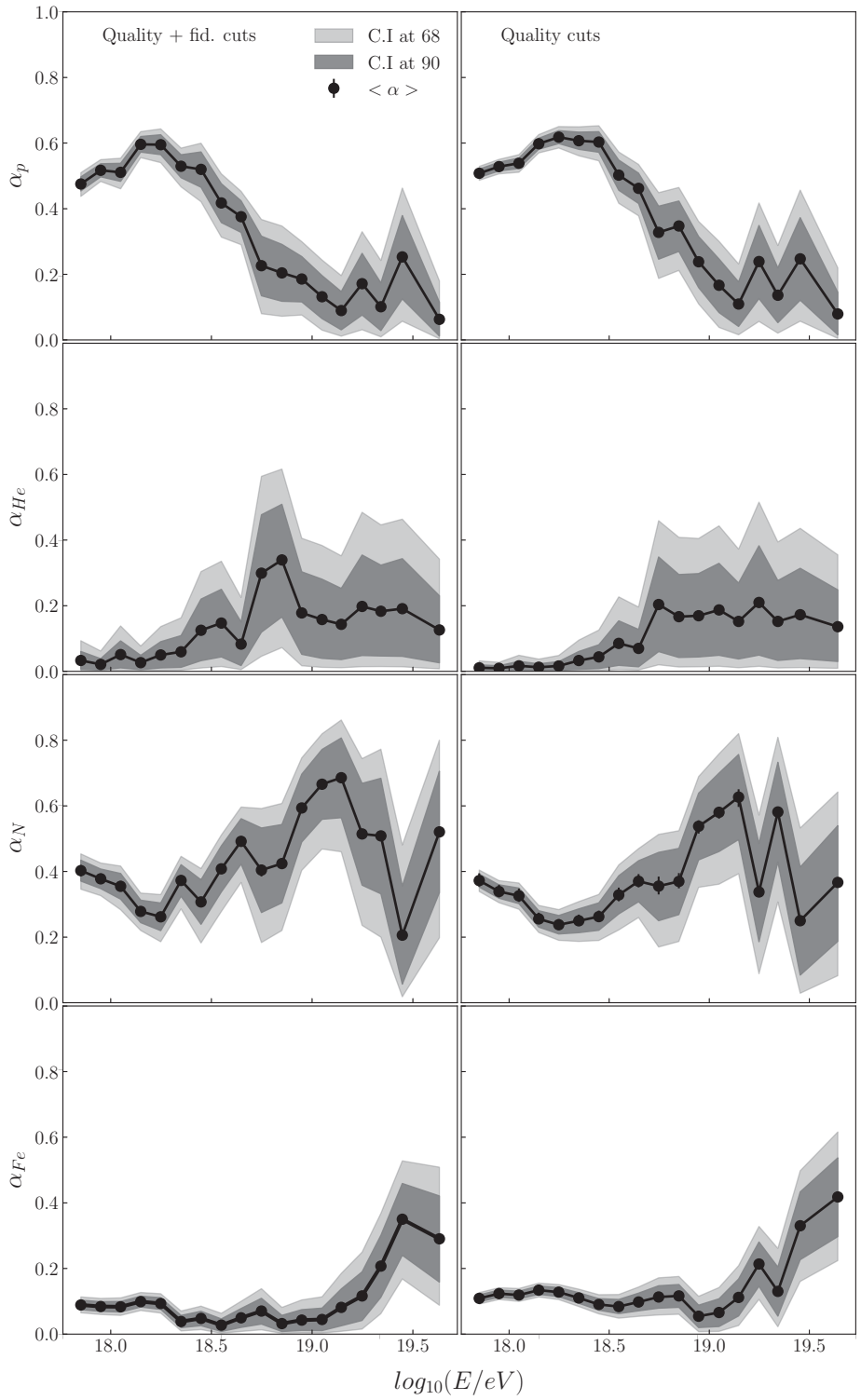


FIGURE 5.21: Composition fractions of p-He-N-Fe from top to bottom scenario using EPOS LHC hadronic interaction model. The analysis using the anti-bias cut is shown in the left column and the analysis without the anti-bias cut is shown in the right column. The mean value (black line) and the 68% and 90% confidence intervals (as bands) are shown in each case.

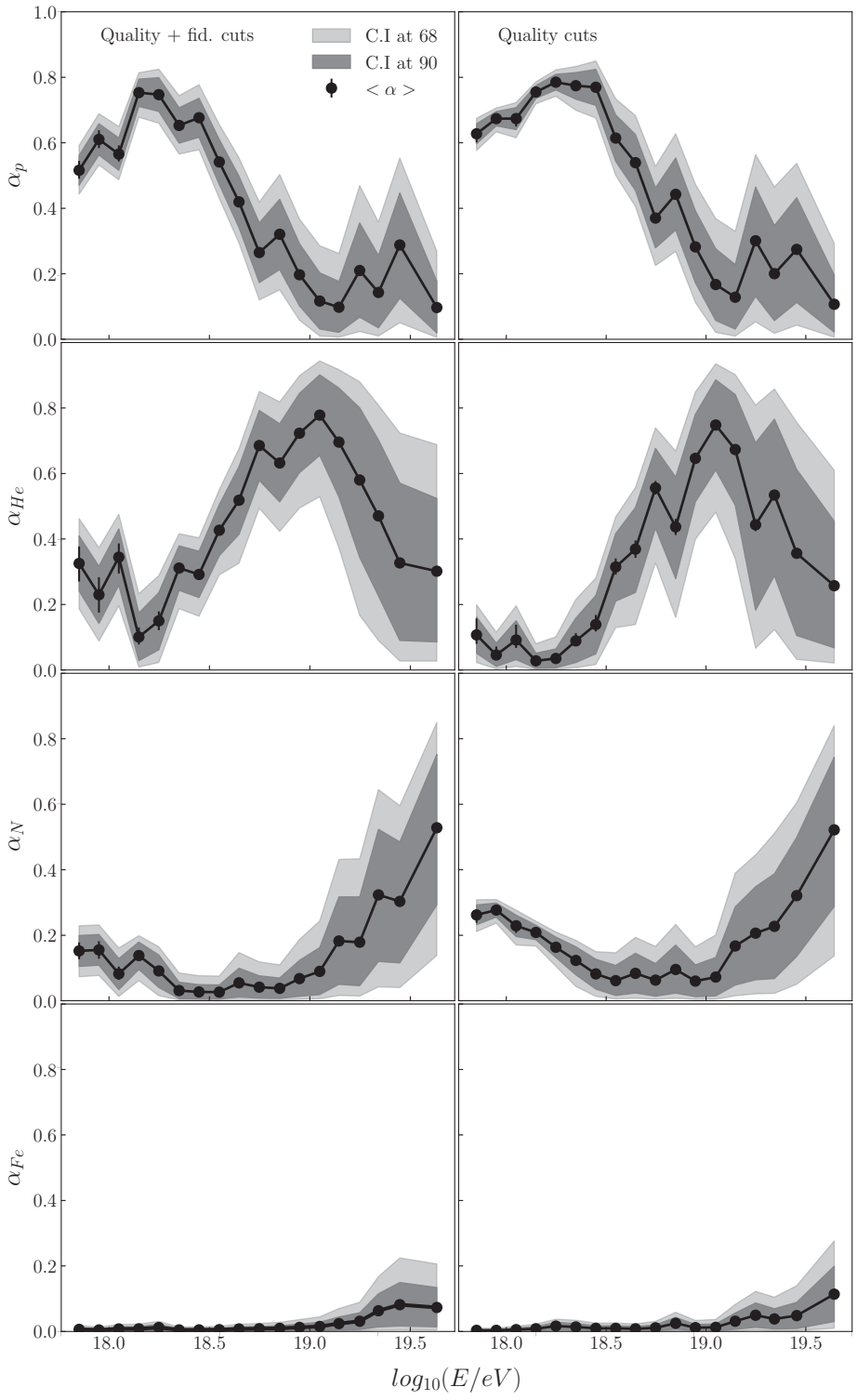


FIGURE 5.22: Same as FIGURE 5.21 but using QGSJETII-04 model.

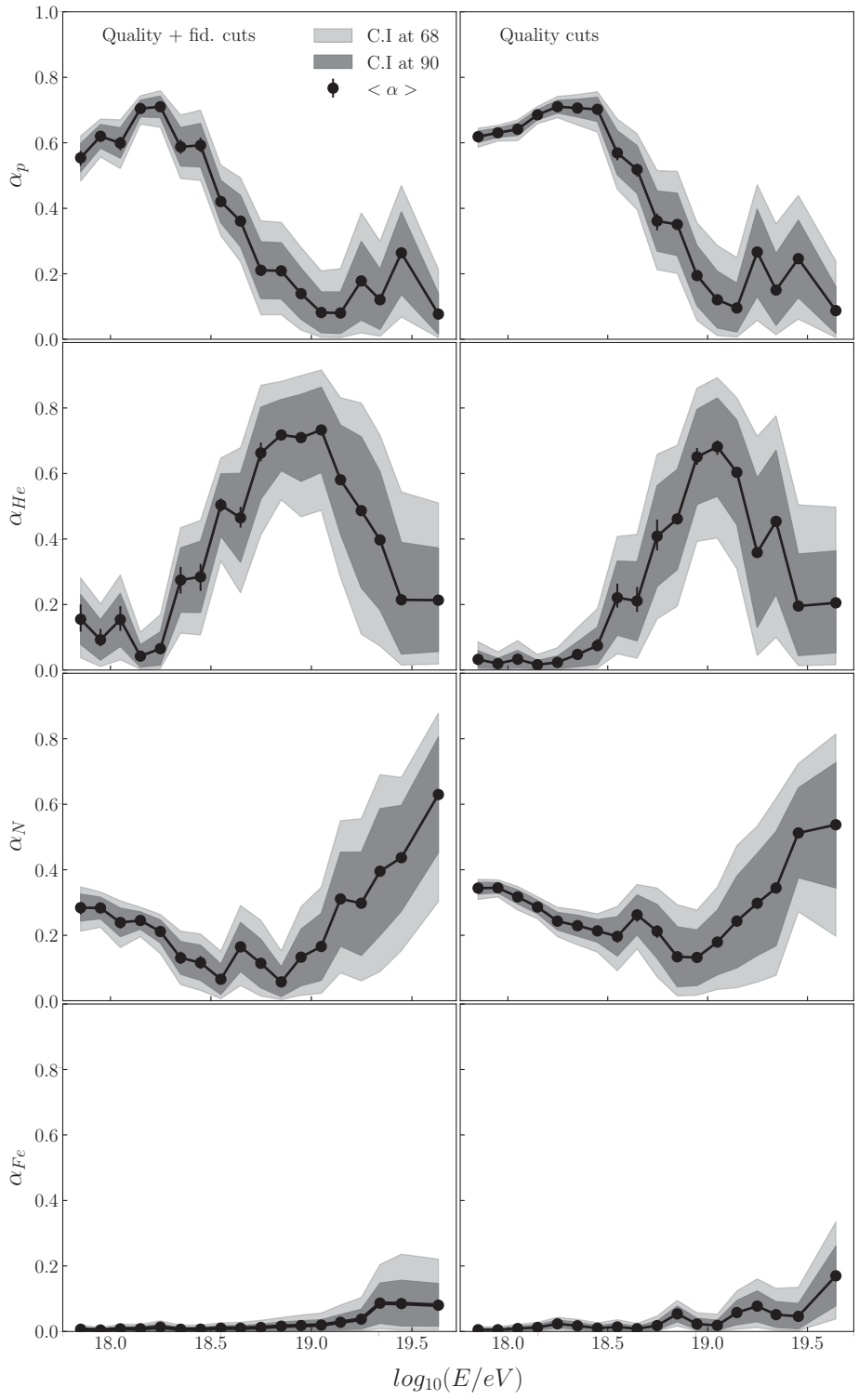


FIGURE 5.23: Same as FIGURE 5.21 but using SIBYLL 2.1 model.

The hypothesis of the Peters' cycle is less favoured when data is analysed using EPOS LHC. When EPOS LHC is used the helium fraction is small and supplied by

more nitrogen and iron than by using SIBYLL 2.1 or QGSJETII-04. In fact, while SIBYLL 2.1 and QGSJETII-04 presents a negligible iron fraction EPOS LHC needs irons to fit the data (with and without fiducial cuts). Of course, the Peters' cycle is not the only hypothesis that can explain the trends of the composition fractions. The behaviour could also be explained assuming that the cosmic rays arriving to the Earth follow a power-law distribution with a shape dependent of the charge.

The comparison of the composition fractions presented in AUGER12 and those obtained in this work are shown in FIGURES 5.24-5.26.

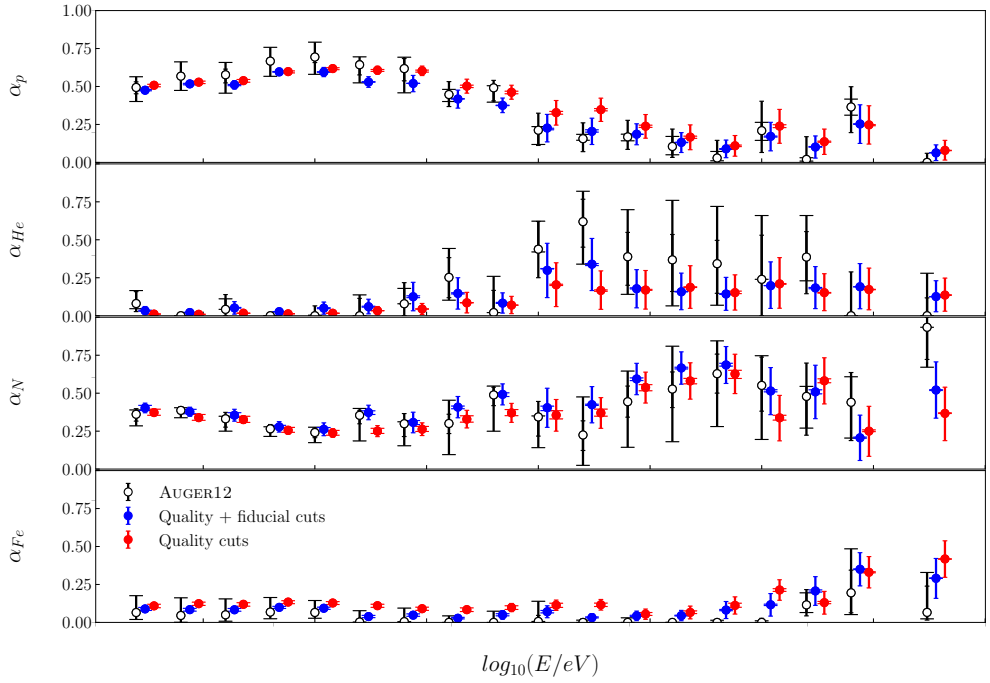


FIGURE 5.24: Comparison of the composition obtained in AUGER12 (white circles) with those obtained using the Bayesian approach with anti-bias (blue squares) and without anti-bias (red squares) applied. The hadronic interaction model used is EPOS LHC.

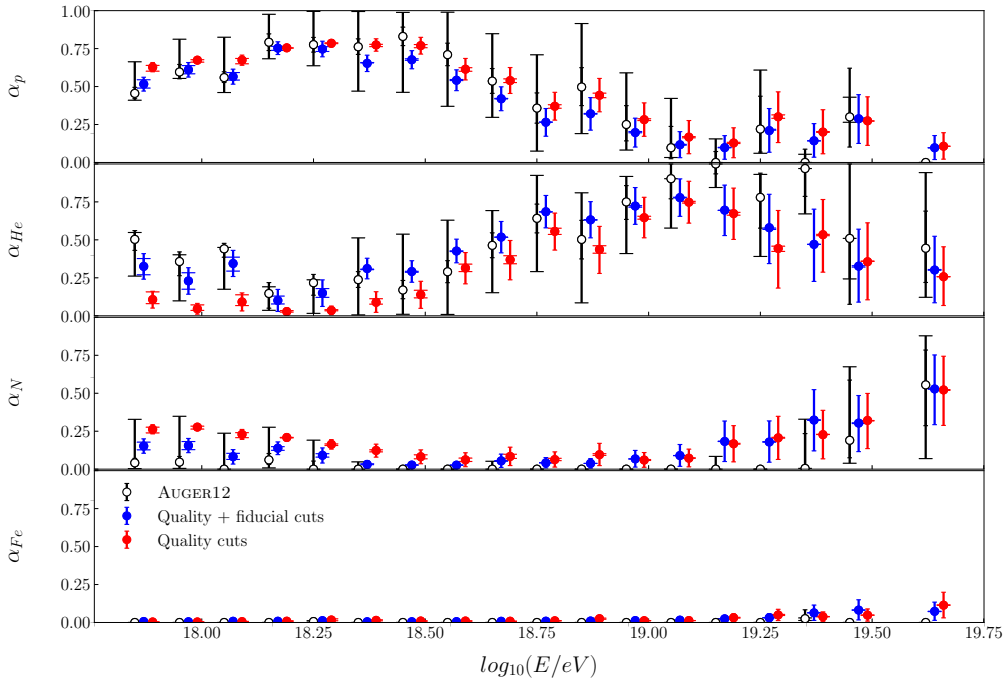


FIGURE 5.25: Same as FIGURE 5.24 but using QGSJETII-04 model.

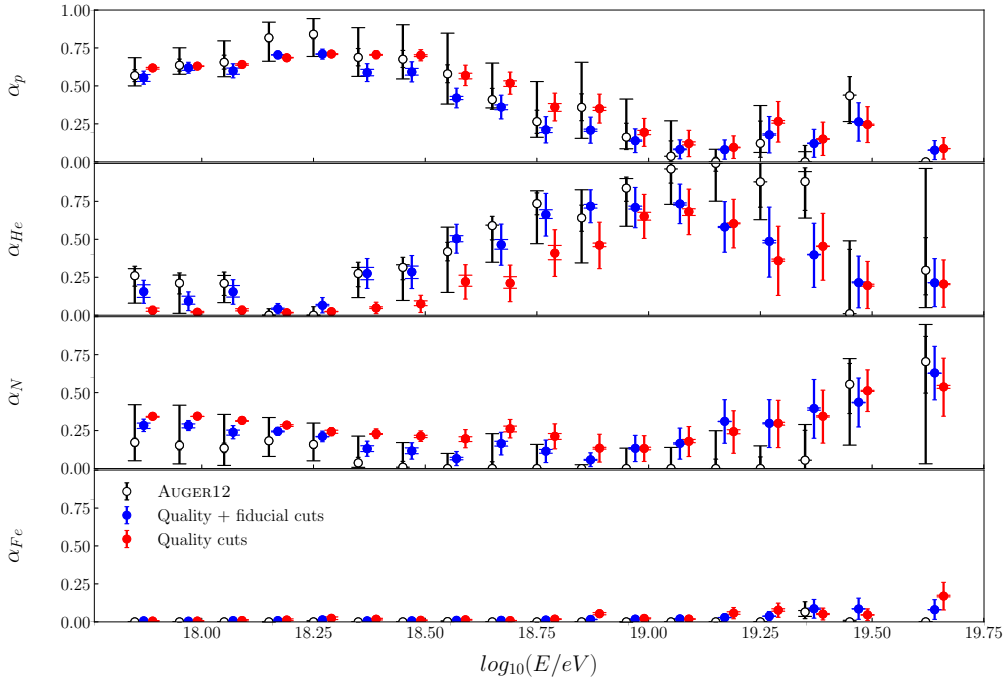


FIGURE 5.26: *Same* as FIGURE 5.24 but using SIBYLL 2.1 model.

The probabilities of the models to describe the data sets are shown in FIGURE 5.27. Once more we note discrepancies between the probabilities using data with fiducial cuts and without fiducial cuts. In the analysis performed using fiducial cuts EPOS LHC is clearly preferred by data (except in $18.1 \leq \log_{10}(E/\text{eV}) < 18.3$ where is QGSJETII-04). Nevertheless, if this cut is removed we find that up to $\log_{10}(E/\text{eV}) = 18.5$ the preferred model is QGSJETII-04 and beyond this energy is EPOS LHC again.

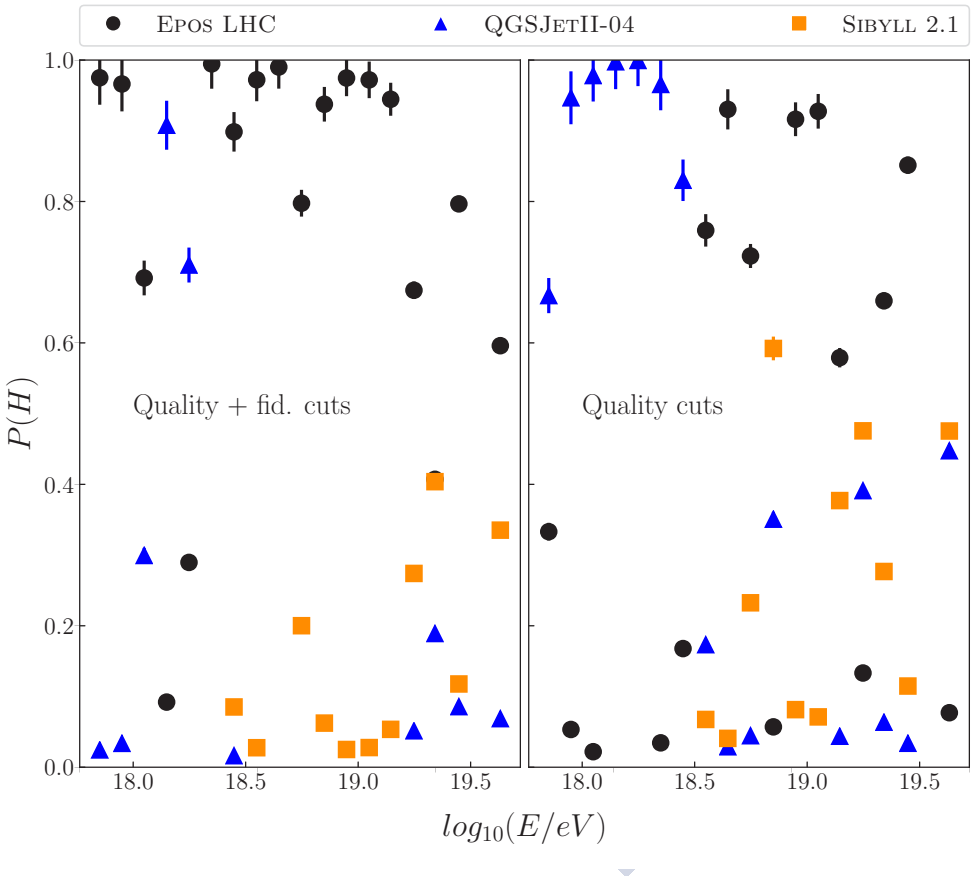


FIGURE 5.27: Probability of the three hadronic interaction models to describe the data at each energy bin.

In this case it is clear that there is a conflict between the data with fiducial cuts and the data without fiducial cuts. Even though this conflict is not so important for the estimated composition, it is for the model comparisons. Removing the anti-bias cut disfavour EPOS LHC. From the point of view of the estimated composition when the anti-bias cut is removed the lightest and heaviest elements increase their relative fractions. An example of this effect is shown if FIGURE 5.28.

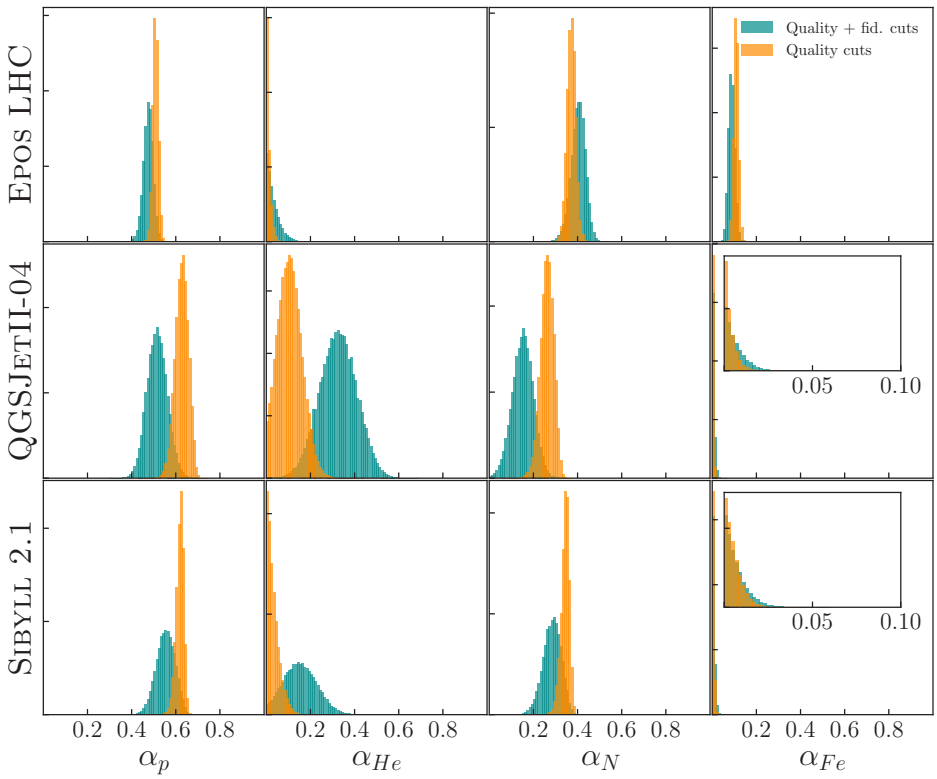


FIGURE 5.28: Marginal posterior p.d.f for proton (first column), helium (second column), nitrogen (third column) and iron (fourth column) using data with anti-bias cut and without anti-bias cut (blue and orange respectively). The analyses performed with the three hadronic interaction models: EPOS LHC (first row), QGSJETII-04 (second row) and SIBYLL 2.1 (third row) are also shown. The energy range is $17.8 < \log_{10}(E/\text{eV}) < 17.9$

In the three models the lightest element found is proton and the analysis performed without fiducial cut increases its fraction. The heaviest elements are iron for EPOS LHC and nitrogen for QGSJETII-04 and SIBYLL 2.1 (iron fraction is practically zero for these two models). For the three models the heaviest element increases its fraction when the fiducial cuts are removed. For the highest energies ($\log_{10}(E/\text{eV}) > 19.5$) the iron fraction is not negligible for QGSJETII-04 and SIBYLL 2.1 and we can see that the iron fraction increases (see FIGURE 5.29).

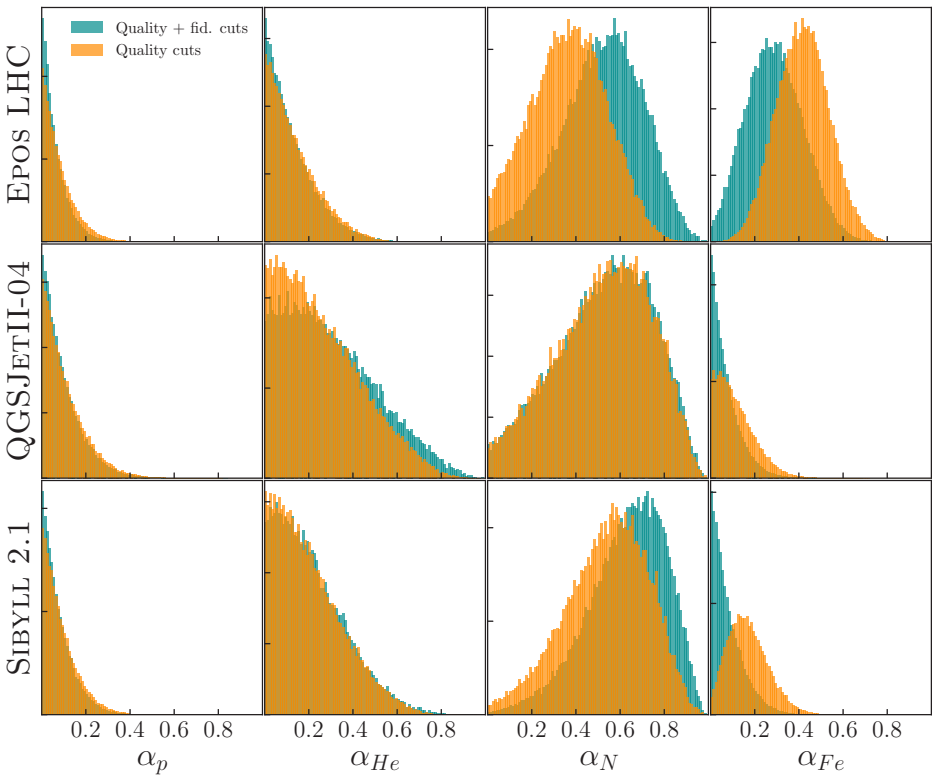
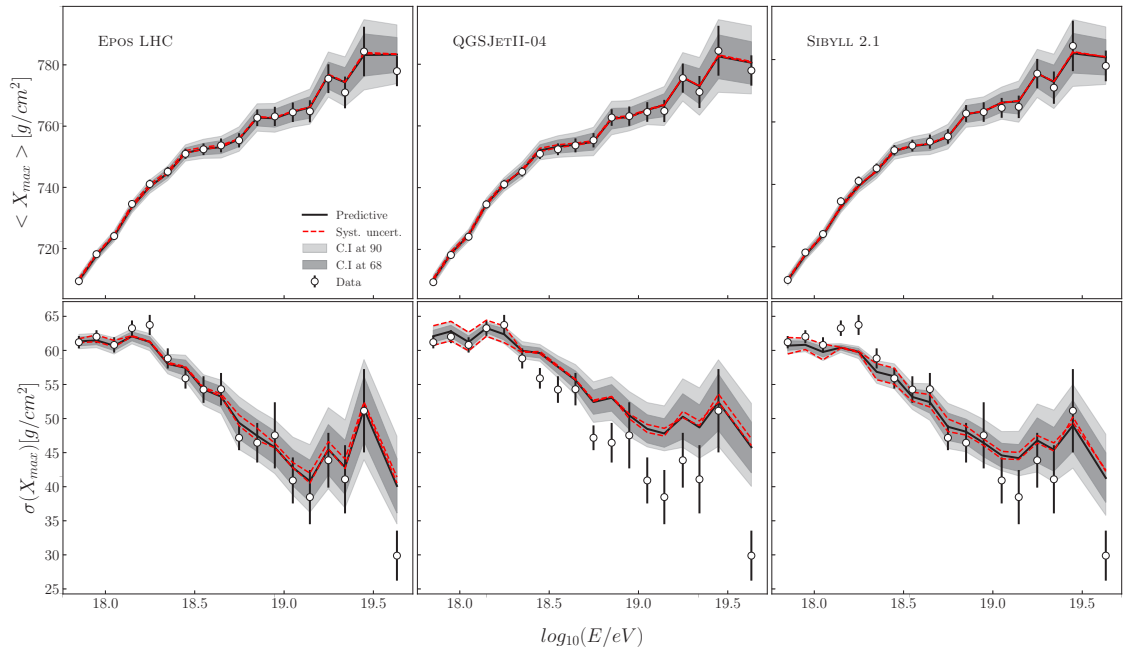
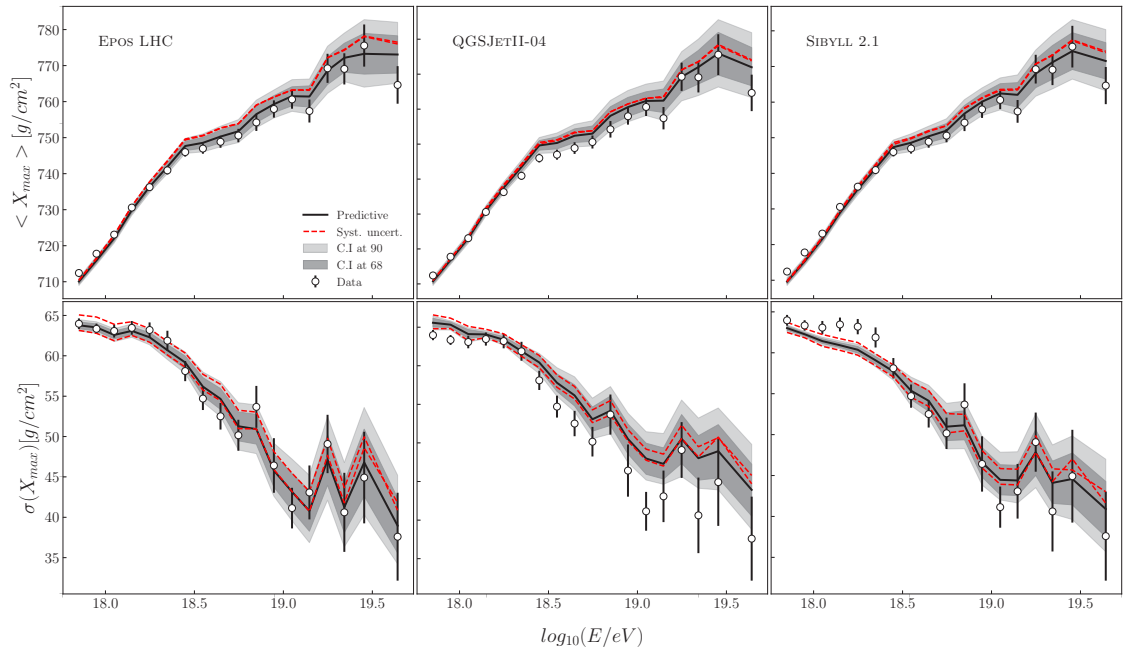


FIGURE 5.29: Same as FIGURE 5.28 but for the last energy bin.

The posterior predictive distributions can be seen in APPENDIX G. In this scenario we show the posterior predictive moments instead the posterior predictive distributions. The posterior predictive moments (average and standard deviation) as a function of the energy can be compared with the measured moments in the same way as the posterior predictive distributions. These comparisons are shown in FIGURE 5.30 for the analysis of events with and without anti-bias cut.



(A)



(B)

FIGURE 5.30: Measured $\langle X_{\max} \rangle$ and $\sigma(X_{\max})$ (upper and lower rows respectively) compared with the posterior predictive moments assuming EPOS LHC (first column), QGSJETII-04 (second column) and SIBYLL 2.1 (third column) for the data set with fiducial cuts (A) and without fiducial cuts (B).

In sight of this figure a conflict between the analysis using anti-bias cut ant without anti-bias cut shows. For both data sets EPOS LHC which predicts better the observed

moments. Nevertheless, when the fiducial cuts are removed at the lowest energies QGSJetII-04 is numerically the preferred model even though the moments are better predicted with Epos LHC.

5.5 p-He-Li-N-Si-Fe scenario

We finally analyse a six-component scenario p-He-Li-N-Si-Fe. The comparison between the estimated composition fractions using different hypotheses can help us to understand the limitations of our measurements due to the resolution of our detector or due to the limited number of events.

It is well know that the abundances lithium, beryllium and boron are suppressed in the Universe. This is due to the production mechanism of these elements during the Big Bang nucleosynthesis and during the burning of elements within the stars. The abundances of the lithium-beryllium-boron group are however larger in low-energy cosmic rays (see [84]). Larger abundances can be explained through the interactions of cosmic rays with the background producing secondary cosmic rays by spallation or photodisintegration. We thus consider lithium as a possible component of cosmic rays arriving to the Earth. The ratio between silicon and nitrogen abundances in the solar system is around 0.1 but here silicon is considered for composition as a group representative keeping a similar distance to N and Fe in log-mass space. Silicon could be also interesting because the separation between the average of $X_{\max}$ distributions between silicon and iron is smaller than between nitrogen and iron. In the previous sections we have seen that at the highest energies there could exist a non-negligible fraction of iron nuclei in the data. This fraction is larger for Epos LHC than for the other models. With the addition of silicon we can check if iron is really needed to describe the data.

The non-overlapping region of the $X_{\max}$ distributions is defined by Equation 3.38. In Figure 5.31 it is shown as a function of energy for this scenario with and without the anti-bias cut. In this figure it is compared with the non-overlapping region of the p-He-N-Fe scenario.

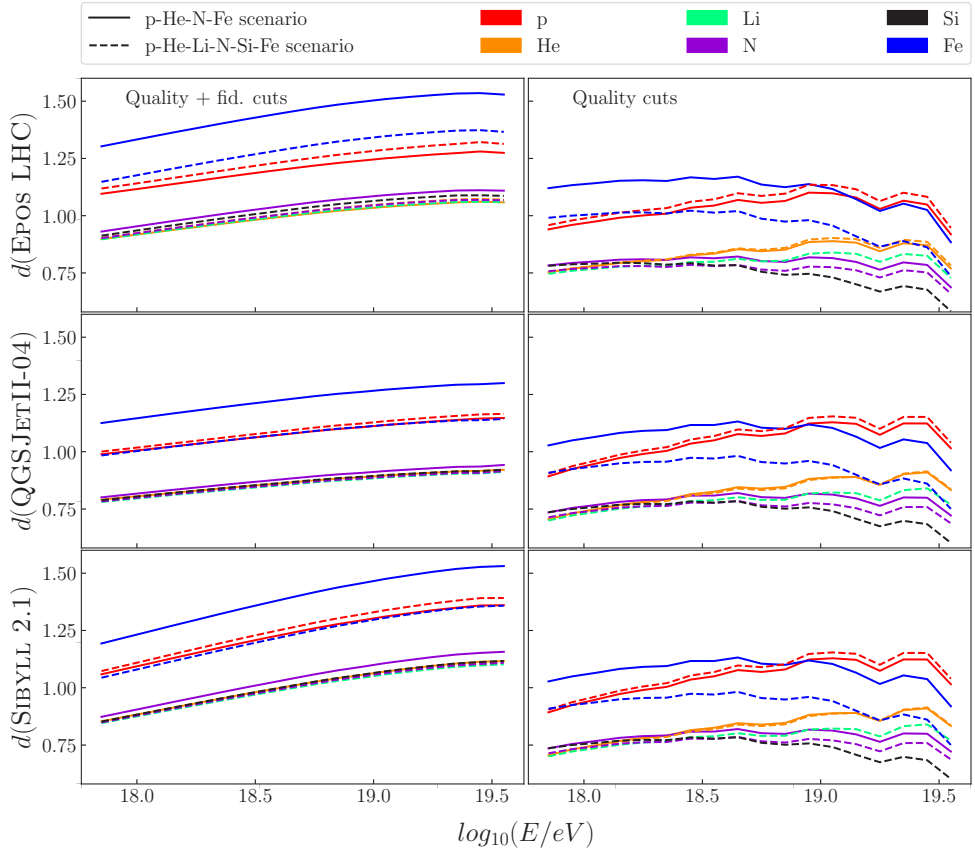


FIGURE 5.31: Distances for the three hadronic models for the distributions applying fiducial cuts and without fiducial cuts for the p-He-N-Fe and p-He-Li-N-Si-Fe scenarios.

By the simple reasoning done in SECTION 3.3.1 we can expect that the estimation of the proton fraction will not be hampered by the addition of lithium and silicon elements to the analysis. This can be seen in TABLE 5.1 where the estimated proton fraction is shown for p-He-N-Fe and p-He-Li-N-Si-Fe scenario for the data set with the fiducial cuts.

	p-He-N-Fe scenario			p-He-Li-N-Si-Fe scenario		
$\Delta\log_{10}(E/\text{eV})$	$\langle\alpha\rangle$	C.I at 68%	C.I at 90%	$\langle\alpha\rangle$	C.I at 68%	C.I at 90%
$[17.8, 17.9)$	0.475	$[0.454, 0.497]$	$[0.437, 0.51]$	0.48	$[0.453, 0.507]$	$[0.433, 0.523]$
$[17.9, 18.0)$	0.517	$[0.497, 0.537]$	$[0.483, 0.55]$	0.538	$[0.514, 0.563]$	$[0.496, 0.578]$
$[18.0, 18.1)$	0.511	$[0.483, 0.538]$	$[0.461, 0.554]$	0.512	$[0.48, 0.545]$	$[0.455, 0.564]$
$[18.1, 18.2)$	0.596	$[0.572, 0.62]$	$[0.556, 0.635]$	0.616	$[0.59, 0.642]$	$[0.571, 0.657]$
$[18.2, 18.3)$	0.595	$[0.565, 0.625]$	$[0.54, 0.643]$	0.596	$[0.562, 0.63]$	$[0.537, 0.65]$
$[18.3, 18.4)$	0.529	$[0.494, 0.564]$	$[0.466, 0.585]$	0.522	$[0.482, 0.561]$	$[0.454, 0.585]$
$[18.4, 18.5)$	0.52	$[0.465, 0.573]$	$[0.421, 0.6]$	0.513	$[0.461, 0.566]$	$[0.422, 0.596]$
$[18.5, 18.6)$	0.418	$[0.358, 0.476]$	$[0.313, 0.506]$	0.396	$[0.34, 0.451]$	$[0.301, 0.485]$
$[18.6, 18.7)$	0.376	$[0.328, 0.424]$	$[0.29, 0.453]$	0.367	$[0.312, 0.421]$	$[0.275, 0.454]$
$[18.7, 18.8)$	0.226	$[0.135, 0.316]$	$[0.08, 0.367]$	0.222	$[0.144, 0.298]$	$[0.095, 0.348]$
$[18.8, 18.9)$	0.204	$[0.118, 0.291]$	$[0.072, 0.347]$	0.193	$[0.122, 0.264]$	$[0.082, 0.315]$
$[18.9, 19.0)$	0.186	$[0.116, 0.254]$	$[0.076, 0.3]$	0.164	$[0.1, 0.228]$	$[0.066, 0.275]$
$[19.0, 19.1)$	0.132	$[0.065, 0.196]$	$[0.03, 0.244]$	0.113	$[0.049, 0.175]$	$[0.022, 0.224]$
$[19.1, 19.2)$	0.09	$[0.031, 0.147]$	$[0.011, 0.196]$	0.08	$[0.024, 0.135]$	$[0.009, 0.185]$
$[19.2, 19.3)$	0.171	$[0.076, 0.264]$	$[0.031, 0.33]$	0.16	$[0.068, 0.251]$	$[0.028, 0.316]$
$[19.3, 19.4)$	0.101	$[0.028, 0.174]$	$[0.009, 0.242]$	0.091	$[0.025, 0.157]$	$[0.008, 0.221]$
$[19.4, 19.5)$	0.253	$[0.125, 0.379]$	$[0.057, 0.463]$	0.212	$[0.103, 0.318]$	$[0.048, 0.395]$
$[19.5, \infty)$	0.063	$[0.012, 0.115]$	$[0.003, 0.18]$	0.057	$[0.012, 0.103]$	$[0.004, 0.16]$

TABLE 5.1: Mean of the proton posterior probability density function and its confidence intervals for different energy ranges for p-He-N-Fe and p-He-Li-N-Si-Fe scenarios.

One can observe that the estimated proton fraction is almost the same for the two primary scenarios in all energy ranges.

Not only the estimated proton fraction is not changed by the addition of the new elements but the confidence intervals are roughly the same as it is illustrated in TABLE 5.2, in this case for both data sets with fiducial and without fiducial cuts.

	p-He-N-Fe scenario				p-He-Li-N-Si-Fe scenario			
	QF		Q		QF		Q	
$\Delta\log_{10}(E/\text{eV})$	Δ_{68}	Δ_{90}	Δ_{68}	Δ_{90}	Δ_{68}	Δ_{90}	Δ_{68}	Δ_{90}
$[17.8, 17.9)$	0.043	0.072	0.025	0.042	0.054	0.09	0.032	0.054
$[17.9, 18.0)$	0.04	0.066	0.027	0.044	0.049	0.082	0.031	0.052
$[18.0, 18.1)$	0.055	0.093	0.032	0.054	0.066	0.109	0.04	0.067
$[18.1, 18.2)$	0.048	0.079	0.034	0.056	0.052	0.086	0.033	0.055
$[18.2, 18.3)$	0.061	0.103	0.039	0.065	0.068	0.113	0.041	0.069
$[18.3, 18.4)$	0.07	0.118	0.052	0.088	0.079	0.132	0.06	0.098
$[18.4, 18.5)$	0.108	0.178	0.063	0.107	0.105	0.174	0.07	0.116
$[18.5, 18.6)$	0.119	0.193	0.092	0.156	0.111	0.185	0.098	0.164
$[18.6, 18.7)$	0.096	0.163	0.092	0.157	0.108	0.179	0.104	0.172
$[18.7, 18.8)$	0.182	0.287	0.162	0.262	0.154	0.253	0.146	0.238
$[18.8, 18.9)$	0.174	0.275	0.154	0.253	0.142	0.233	0.146	0.239
$[18.9, 19.0)$	0.138	0.224	0.156	0.251	0.129	0.209	0.151	0.247
$[19.0, 19.1)$	0.131	0.214	0.163	0.263	0.125	0.202	0.157	0.251
$[19.1, 19.2)$	0.117	0.185	0.136	0.216	0.111	0.176	0.13	0.207
$[19.2, 19.3)$	0.188	0.299	0.223	0.361	0.183	0.288	0.206	0.333
$[19.3, 19.4)$	0.145	0.233	0.167	0.267	0.132	0.213	0.162	0.258
$[19.4, 19.5)$	0.254	0.406	0.252	0.4	0.215	0.347	0.21	0.338
$[19.5, \infty)$	0.103	0.177	0.13	0.215	0.092	0.156	0.106	0.178

TABLE 5.2: Differences between the upper and lower values of the confidence intervals at 68% and 90% denoted as Δ_{68} and Δ_{90} respectively for each energy bin. These values are shown for p-He-N-Fe and p-He-Li-N-Si-Fe scenarios and for data with anti-bias cut (QF) and without anti-bias cut (Q).

The inferences of the six primaries are shown in FIGURES 5.32-5.34 for the three hadronic interaction models.

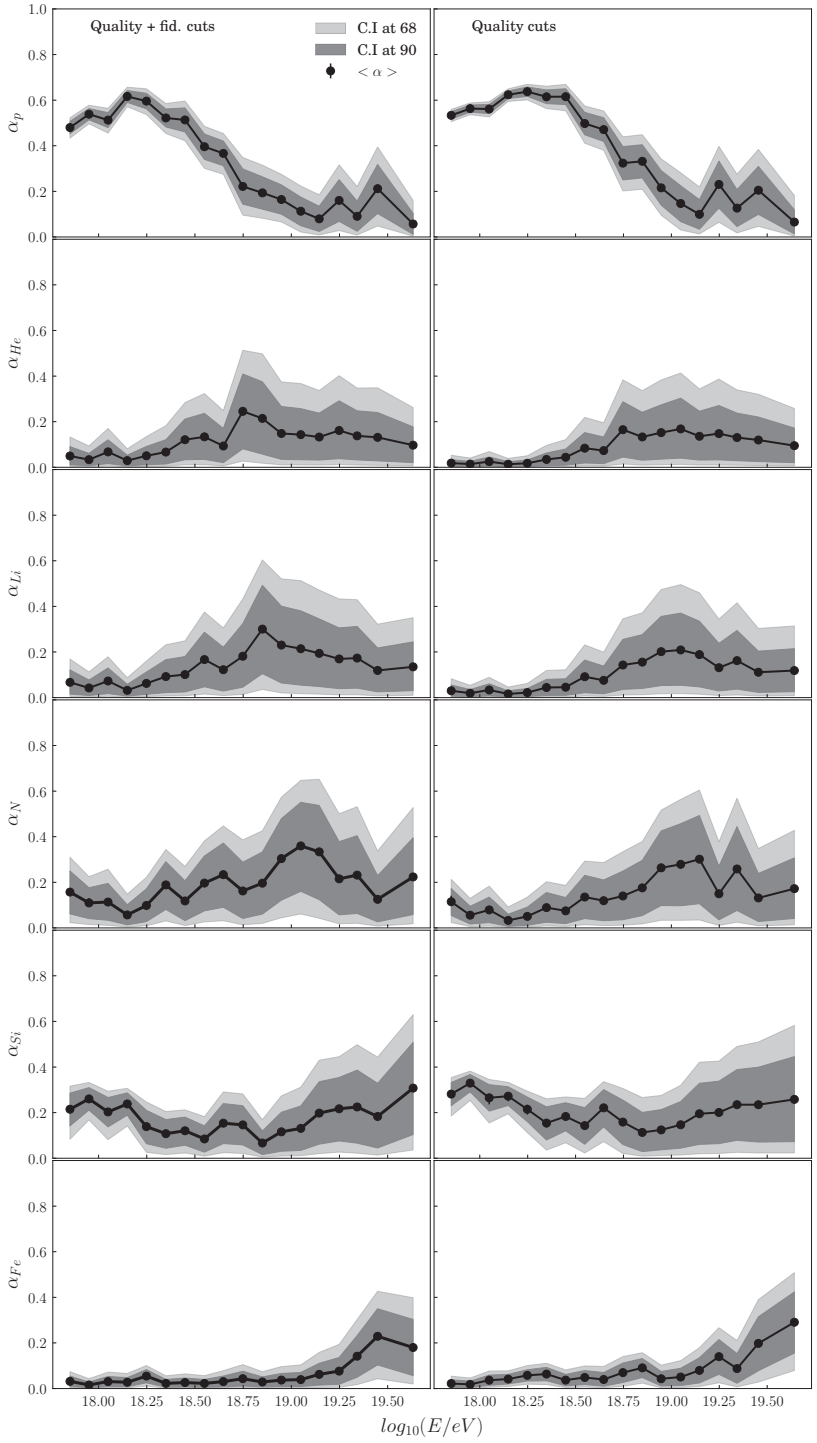


FIGURE 5.32: Composition fractions of p, He, Li, N, Si and Fe from top to bottom using EPOS LHC. The analysis using (not using) the anti-bias cut is shown in the left (right) column. For all primaries the mean value and the 68% and 90% confidence intervals are shown together.

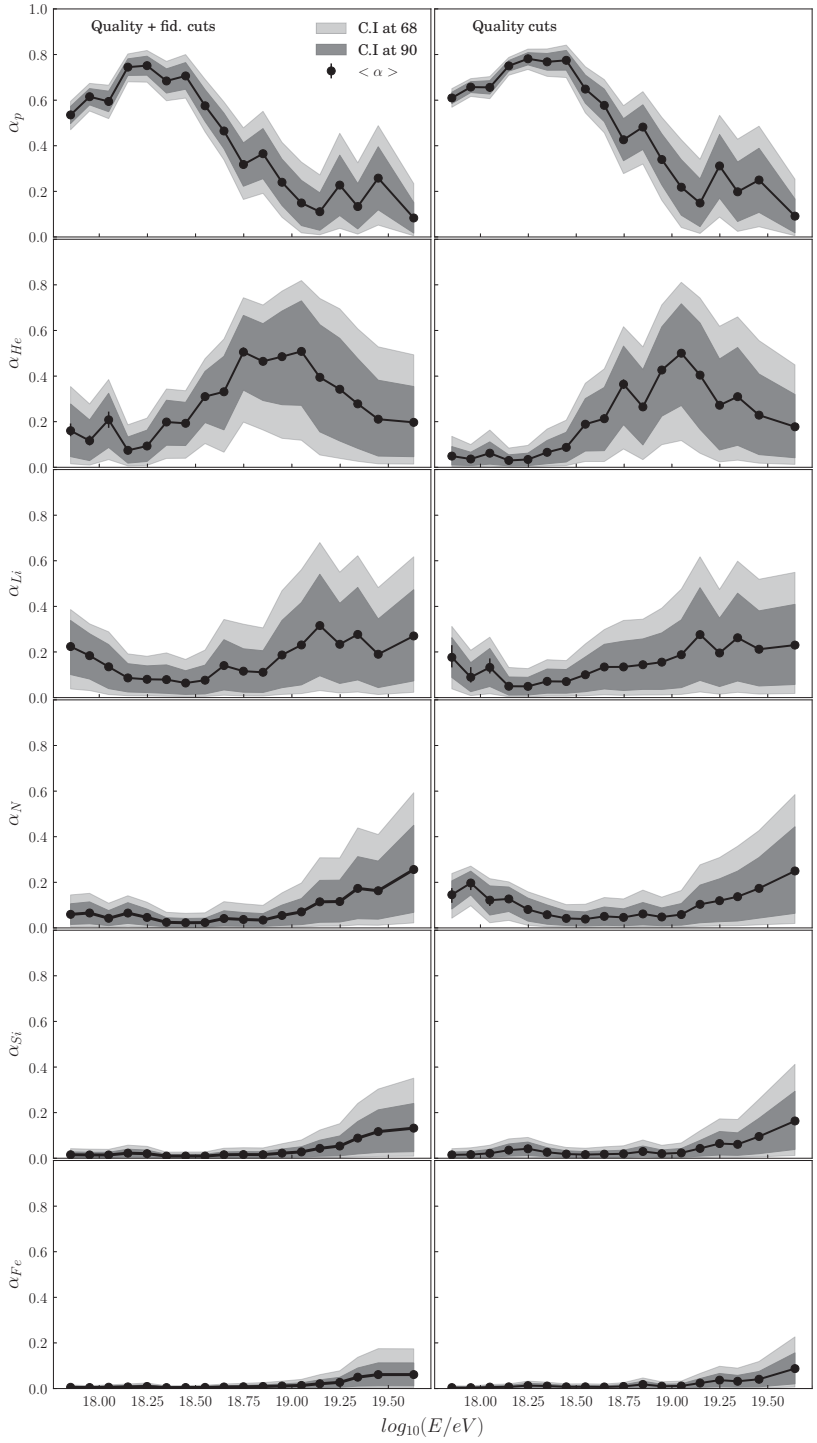


FIGURE 5.33: Same as FIGURE 5.32 but using QGSJETII-04 model.

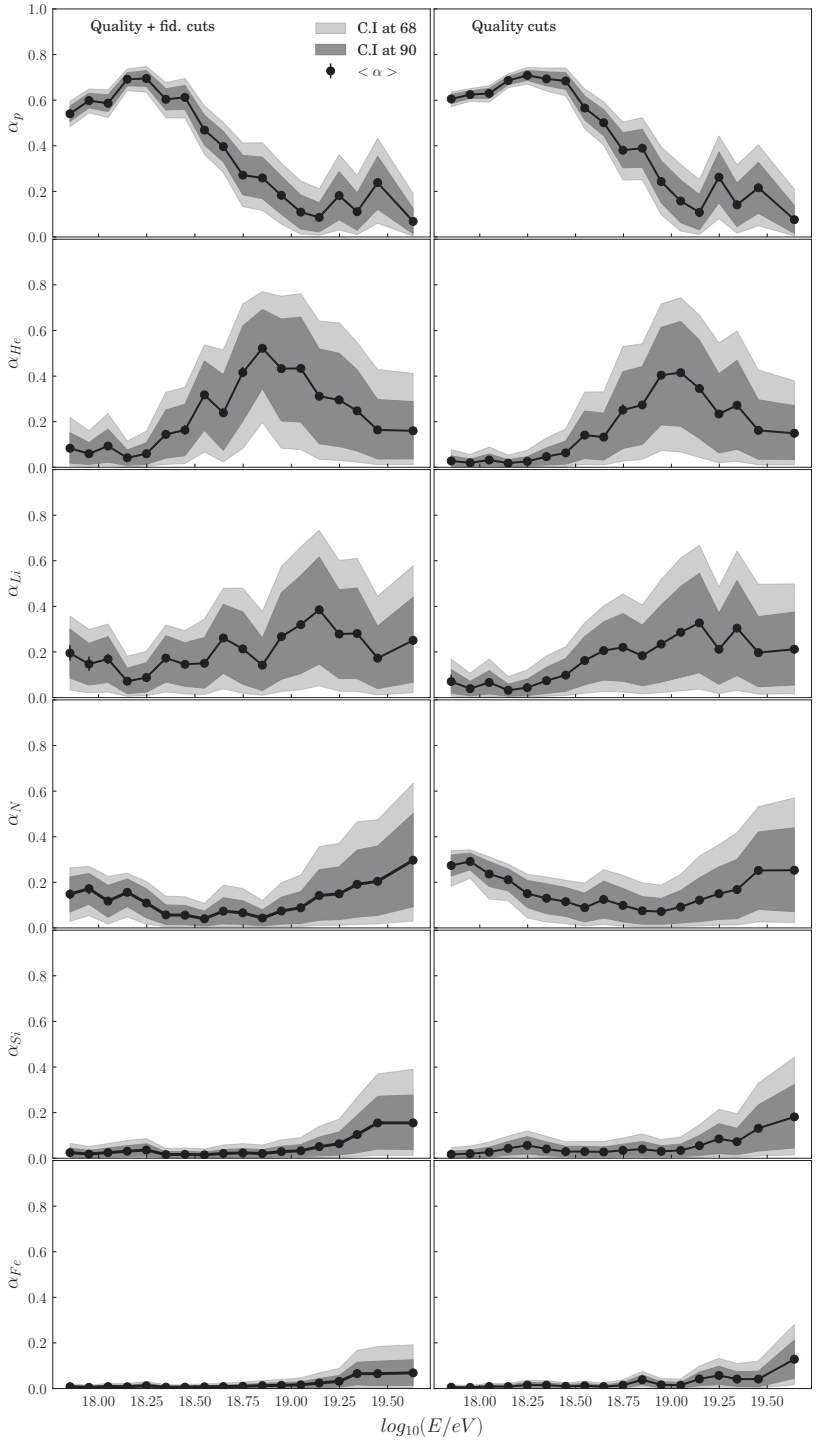


FIGURE 5.34: Same as FIGURE 5.32 but using SIBYLL 2.1 model.

As discussed for the four primary scenario, the protons dominate over the other primaries at the lowest energies. Its fraction increases reaching a maximum and then it drops. This behaviour is still fulfilled by the three hadronic interaction models.

Observing the inferences done using EPOS LHC the most remarkable results passing from the four-primary scenario to the six-primary scenario are: the reduction of the nitrogen fraction in favour of silicon and the disappearance of the iron fraction at lowest energies. At higher energies the iron fraction also appears in the p-He-Li-N-Si-Fe scenario although its fraction is lower than in the four-primary scenario. Events that were associated to nitrogen and iron in the p-He-N-Fe scenario are now associated to silicon. Some numerical comparisons between the four-primary and six-primary scenarios are shown in TABLE 5.3 for three selected energy bins.

	$18 \leq \log_{10}(E/\text{eV}) < 18.1$	$19 \leq \log_{10}(E/\text{eV}) < 19.1$	$19.5 \leq \log_{10}(E/\text{eV}) < \infty$
$\langle\alpha_p\rangle$	$(0.51, 0.51)$	$(0.13, 0.11)$	$(0.06, 0.06)$
$\langle\alpha_{He}\rangle$	$(0.05, 0.07)$	$(0.16, 0.14)$	$(0.13, 0.1)$
$\langle\alpha_{Li}\rangle$	$(-, 0.07)$	$(-, 0.21)$	$(-, 0.13)$
$\langle\alpha_N\rangle$	$(0.35, 0.11)$	$(0.67, 0.36)$	$(0.52, 0.22)$
$\langle\alpha_{Si}\rangle$	$(-, 0.2)$	$(-, 0.13)$	$(-, 0.31)$
$\langle\alpha_{Fe}\rangle$	$(0.08, 0.03)$	$(0.04, 0.04)$	$(0.29, 0.18)$

TABLE 5.3: Comparison of the mean of posterior p.d.f of fractions for three energy ranges. At each energy column the first number represents the estimation in the p-He-N-Fe scenario and the second is the estimation in the p-He-Li-N-Si-Fe scenario. Lithium and silicon are not considered in the four-primary scenario and are represented by "–". The inferences are obtained using EPOS LHC hadronic interaction model and data passing both quality and fiducial cuts.

These changes are similar when QGSJETII-04 and SIBYLL 2.1 are used. The fractions of helium and nitrogen are reduced when we add lithium and silicon primaries. Since the iron fraction is small using these hadronic interaction models, its inference does not change due to the presence of the new elements. Numerical comparisons between p-He-N-Fe and p-He-Li-N-Si-Fe scenarios for some energy ranges are also shown in TABLES 5.4-5.5 for QGSJETII-04 and SIBYLL 2.1 respectively.

	$18 \leq \log_{10}(E/\text{eV}) < 18.1$	$19 \leq \log_{10}(E/\text{eV}) < 19.1$	$19.5 \leq \log_{10}(E/\text{eV}) < \infty$
$\langle \alpha_p \rangle$	$(0.57, 0.59)$	$(0.12, 0.15)$	$(0.1, 0.08)$
$\langle \alpha_{He} \rangle$	$(0.34, 0.21)$	$(0.78, 0.51)$	$(0.3, 0.2)$
$\langle \alpha_{Li} \rangle$	$(-, 0.14)$	$(-, 0.23)$	$(-, 0.27)$
$\langle \alpha_N \rangle$	$(0.08, 0.04)$	$(0.09, 0.07)$	$(0.53, 0.26)$
$\langle \alpha_{Si} \rangle$	$(-, 0.01)$	$(-, 0.03)$	$(-, 0.13)$
$\langle \alpha_{Fe} \rangle$	$(0.01, 0.01)$	$(0.02, 0.01)$	$(0.07, 0.06)$

TABLE 5.4: Same as TABLE 5.3 but using QGSJETII-04 hadronic model.

	$18 \leq \log_{10}(E/\text{eV}) < 18.1$	$19 \leq \log_{10}(E/\text{eV}) < 19.1$	$19.5 \leq \log_{10}(E/\text{eV}) < \infty$
$\langle \alpha_p \rangle$	$(0.6, 0.59)$	$(0.08, 0.11)$	$(0.08, 0.07)$
$\langle \alpha_{He} \rangle$	$(0.15, 0.09)$	$(0.73, 0.43)$	$(0.21, 0.16)$
$\langle \alpha_{Li} \rangle$	$(-, 0.17)$	$(-, 0.32)$	$(-, 0.25)$
$\langle \alpha_N \rangle$	$(0.24, 0.12)$	$(0.17, 0.09)$	$(0.63, 0.3)$
$\langle \alpha_{Si} \rangle$	$(-, 0.02)$	$(-, 0.03)$	$(-, 0.15)$
$\langle \alpha_{Fe} \rangle$	$(0.01, 0.01)$	$(0.02, 0.02)$	$(0.08, 0.07)$

TABLE 5.5: Same as TABLE 5.3 but using SIBYLL 2.1 hadronic model.

The probabilities of the models are shown in FIGURE 5.35. One can observe that in this scenario EPOS LHC model is clearly the preferred model over the other two in both data samples with and without fiducial cuts.

There is a remarkable difference between FIGURE 5.35 and FIGURE 5.27. In FIGURE 5.27 (p-He-N-Fe scenario) the most probable model is EPOS LHC when the data set with fiducial cuts is used but when the fiducial cuts are removed QGSJETII-04 becomes the preferred model at lower energies. In the six-primary scenario is EPOS LHC the preferred model in all the energy bins when the fiducial cuts are applied and also in almost all energy bins when the fiducial cuts are removed. The discrimination among models when data with and without fiducial cuts are compared gives more consistent results in the six-primary scenario than in the four-primary scenario. The expected value and standard deviation of $X_{\max}$ using the posterior predictive distributions are displayed in FIGURE 5.36 and we can see that EPOS LHC reproduce these moments much better than the other models.

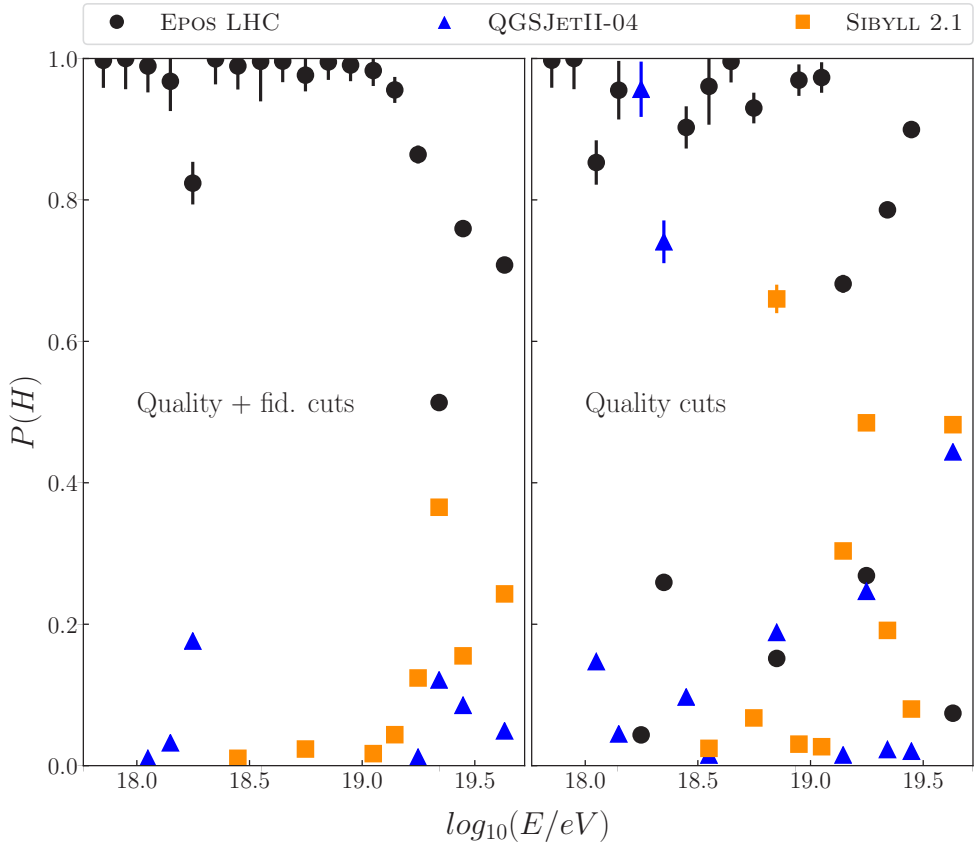
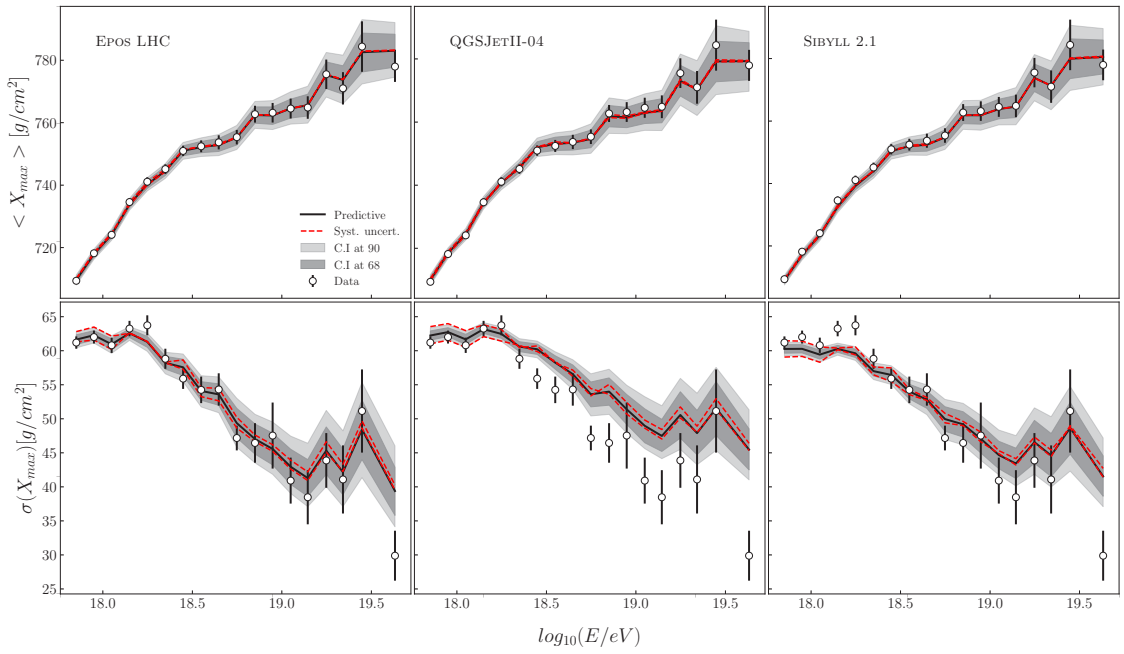
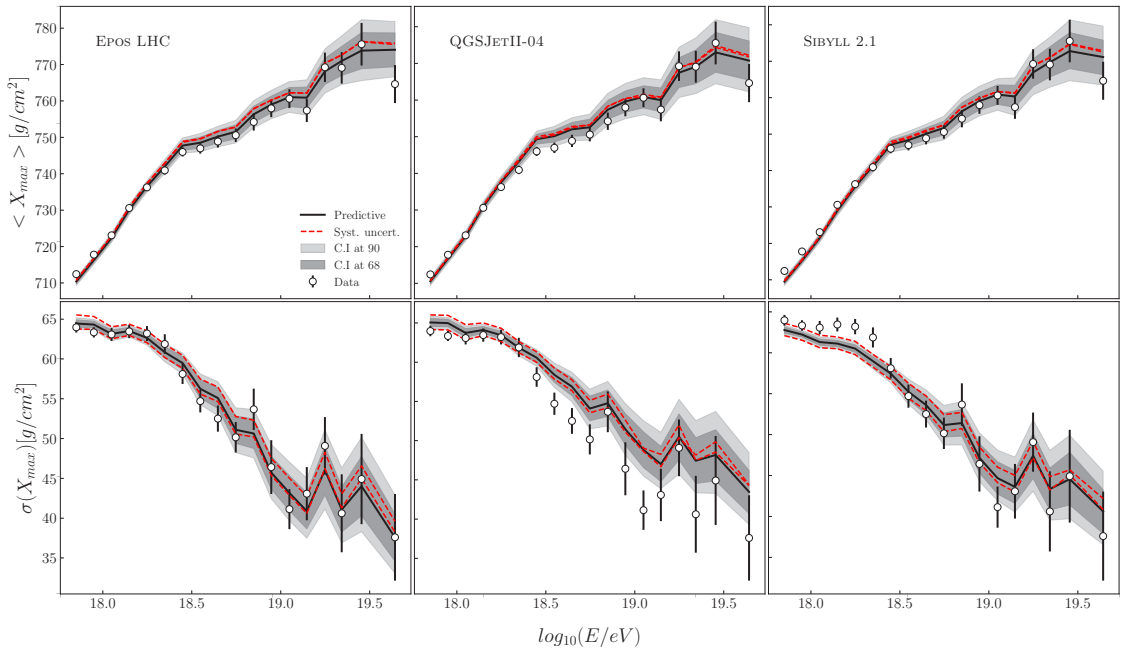


Figure 5.35: Probability of the three hadronic interaction models: Epos LHC (black circles), QGSJetII-04 (blue triangles) and Sibyll 2.1 (orange squares). The probabilities using data with quality and fiducial cuts are shown in the left panel. The right panel is for data without fiducial cuts.

The differences between the four-primary and six-primary scenarios in the analysis when the fiducial cuts are removed could be due to the assumption about the number of primaries. If we assume that only p-He-N-Fe arrive to the Earth but more elements are actually in the data, the expected bias in composition that we take into account when we perform the analysis without fiducial cuts is badly estimated. As we add more primaries we can calculate more accurately the bias in composition that exists in data. Probably there are more elements of elements between proton and iron or whatever is the highest element reaching our detector. If we assume that the hadronic models are correct ant the characterisation of our detector is complete, as more elements are included in our analysis inference becomes more accurate. Alternative our description of the detector could be incomplete. We already know that the models are not completely satisfactory to explain all the data from the Observatory ([36] and [45]). Establishing the final cause of this behaviour will require further work.



(A)



(B)

FIGURE 5.36: Posterior predictive mean and standard deviations of $X_{\max}$ as a function of energy for events with quality and fiducial cuts (A) and events with only quality cuts (B).

5.6 Analysis of the scenarios I

We have observed that removing the fiducial does not introduce significant changes in the estimation of the composition. However the conclusions about different models are different when these cuts are removed. The rest of these chapter will be done using the fiducial cuts so the results can be compared with the official results of the Pierre Auger Observatory. Nevertheless, the plots showing the composition fractions will contain the inferences without fiducial cuts for comparison.

Scenarios and hadronic models can be compared to each other. The best scenario for each model together with the best model are shown in Table 5.6 for all energy ranges.

$\Delta\log_{10}(E/\text{eV})$	Epos LHC	QGSJetII-04	Sibyll 2.1	Best Model
$[17.8, 17.9)$	p-N-Fe	p-He-N	p-He-N	Epos LHC
$[17.9, 18)$	p-He-Li-N-Si-Fe	p-He-N	p-N	Epos LHC
$[18, 18.1)$	p-He-Li-N-Si-Fe	p-He-N	p-N	QGSJetII-04
$[18.1, 18.2)$	p-He-Li-N-Si-Fe	p-N	p-N	QGSJetII-04
$[18.2, 18.3)$	p-N-Fe	p-N	p-N	QGSJetII-04
$[18.3, 18.4)$	p-N	p-He	p-He-N	Epos LHC
$[18.4, 18.5)$	p-N	p-He	p-He-N	QGSJetII-04
$[18.5, 18.6)$	p-N	p-He	p-He	Epos LHC
$[18.6, 18.7)$	p-N	p-He	p-He-N	Epos LHC
$[18.7, 18.8)$	p-He-Li-N-Si-Fe	p-He	p-He-N	Sibyll 2.1
$[18.8, 18.9)$	p-He-N	p-He	p-He	Epos LHC
$[18.9, 19)$	p-N	p-He	p-He-N	Epos LHC
$[19, 19.1)$	p-N	p-He	He-N	Epos LHC
$[19.1, 19.2)$	p-N	He-N	He-N	Epos LHC
$[19.2, 19.3)$	p-He-Li-N-Si-Fe	p-He	He-N	Sibyll 2.1
$[19.3, 19.4)$	p-He-Li-N-Si-Fe	He-N	He-N	Sibyll 2.1
$[19.4, 19.5)$	p-N-Fe	p-He-N	p-N	Epos LHC
$[19.5, \infty)$	N-Fe	He-N	He-N	Epos LHC

Table 5.6: Best composition scenarios for each energy bin and hadronic model: Epos LHC (first column), QGSJetII-04 (second column) and Sibyll 2.1 (third column). The best hadronic model is shown in the last column.

One of the most remarkable characteristics using Bayesian inference to select the preferred model is that simpler model is preferred against more complex model if both models can describe the observed data (see APPENDIX D). For this reason a scenario with less number of primaries should be preferred over another scenario with more number of primaries if both scenarios can reproduce equivalently the data.

To interpret the results of TABLE 5.6 we are going to take a look at the inference in the $18 \leq \log_{10}(E/\text{eV}) < 18.1$ energy range. If we analyse the data using SIBYLL 2.1 the preferred scenario is p-N with a proton fraction $\alpha_p = 0.68^{+0.03}_{-0.02}$ and a nitrogen fraction $\alpha_N = 0.32^{+0.01}_{-0.02}$. We compare this estimations with other scenarios in TABLE 5.7 using the same hadronic interaction model.

Scenario	$\langle\alpha\rangle$	$\mathcal{H}$	Probability	$\ln\mathcal{L}(\langle\alpha\rangle)$
p-N	$(0.68, 0.32)$	2.66	0.5	-15384.84
p-He-N	$(0.61, 0.13, 0.26)$	3.29	0.48	-15383.91
p-He-N-Fe	$(0.6, 0.15, 0.24, 0.01)$	5.93	0.01	-15384.87
p-N-Fe	$(0.68, 0.31, 0.)$	6.25	0.01	-15385.41
p-He-Li-N-Si-Fe	$(0.59, 0.09, 0.17, 0.12, 0.02, 0.01)$	7.11	≈ 0	-15384.7

TABLE 5.7: Mean values of the posterior p.d.f, information gain, probability and log-likelihood for the mean value of the posterior p.d.f for different primary scenarios (the five scenarios shown are the most probable). The data sample analysed belongs to the energy range $18 \leq \log_{10}(E/\text{eV}) < 18.1$ and SIBYLL 2.1 is the hadronic interaction model used for the analysis.

One can observe that in this energy range both p-N and p-He-N scenarios have almost the same probability to describe the data sample. Moreover the likelihood is larger for the p-He-N scenario than for the p-N scenario but the Bayesian model comparison takes into account the simplicity of the model. Note that in the results presented in TABLE 5.7 we assume that all primary scenarios are *a priori* equally probable. In this work we assume this prior for the models but the results could change assuming other priors related with the probability of observing nitrogen but not helium, etc. The study of other priors is out of the scope of this work but it is interesting to study of other choices of priors. Thee could be related with astronomical assumptions such as the probability of observing a given element in relation to another. Such a study is would need to account for the change in probability of observing given elements in relation to the distance to the sources, the magnetic fields that would affect the transport of the cosmic rays and the acceleration mechanisms.

It is also interesting to observe that the worst of the scenarios presented in TABLE 5.7 has also a larger likelihood than the preferred model but note that the the space of parameters is much more complicated having six parameters to fit instead only two. The ratio between the "volume" of the parameter space of p-He-Li-N-Si-Fe scenario and p-N scenario is $Beta((1,1,1,1,1,1))/Beta((1,1)) = 1/120$. That means that the *a priori* ratios of probabilities favours the p-N scenario 120 times more than the p-He-Li-N-Si-Fe scenario.

One can also observe that the gain of information increases as the number of parameters increases. That is again an effect of the volume of the prior. Remember that the information gain is a measure of how much peaked the likelihood is, which is related to the ratio between the posterior volume and the prior volume.

The comparison among posterior predictive distributions and the observed data for the best three scenarios are shown in FIGURE 5.37. Note that the posterior predictive of the p-He-N and p-He-N-Fe scenarios are almost the same but p-He-N is much preferred than the p-He-N-Fe scenario. The explanation of this fact is again the same, *i.e.*, the simpler scenario, and understood by looking the estimated composition fraction. The estimated fractions are $\langle\alpha\rangle_{p-He-N} = (0.61, 0.13, 0.26)$ and $\langle\alpha\rangle_{p-He-N-Fe} = (0.6, 0.15, 0.24, 0.01)$ for p-He-N and p-He-N-Fe respectively. The estimated fractions are almost the same and the differences between probabilities are telling us that iron is not really needed to describe the data.

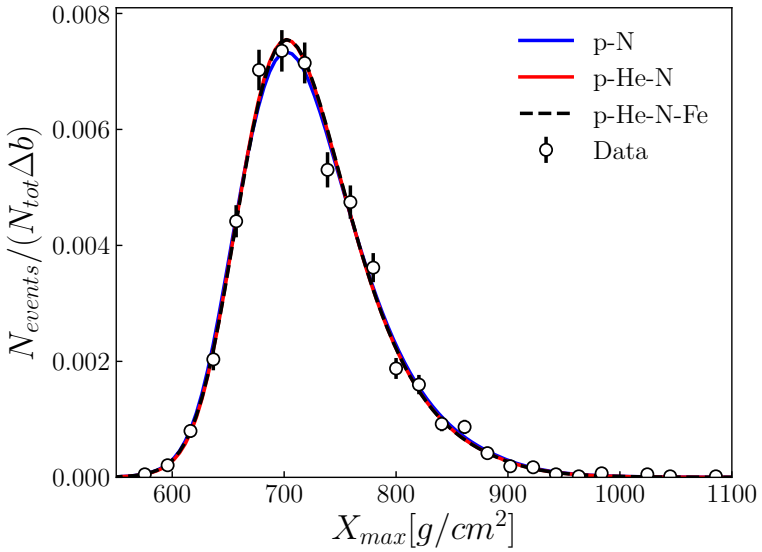


FIGURE 5.37: Comparison of the $X_{\max}$ data histogram at energies $18 \leq \log_{10}(E/\text{eV}) < 18.1$ and the posterior predictive distributions for the best three scenarios using SIBYLL 2.1 hadronic interaction model.

The five best primary scenarios with their respective probabilities, estimated fractions, likelihoods and information gains are shown in TABLE 5.8 for QGSJETII-04 and TABLE 5.9 for EPOS LHC in the energy range $18 \leq \log_{10}(E/\text{eV}) < 18.1$. The comparison between the data distribution and the posterior predictive distribution for each of the three best primary scenarios of the different hadronic models are shown in FIGURE 5.38 and 5.39 respectively for QGSJETII-04 and EPOS LHC.

Scenario	$\langle\alpha\rangle$	$\mathcal{H}$	Probability	$\ln\mathcal{L}(\langle\alpha\rangle)$
p-He-N	$(0.57, 0.32, 0.11)$	2.9	0.71	-15380.58
p-He	$(0.48, 0.52)$	1.97	0.24	-15383.03
p-He-Fe	$(0.51, 0.48, 0.01)$	4.61	0.02	-15382.47
p-He-N-Fe	$(0.57, 0.34, 0.08, 0.01)$	5.83	0.02	-15381.47
p-N	$(0.74, 0.26)$	2.52	≈ 0	-15386.59

TABLE 5.8: Same as TABLE 5.7 but using QGSJETII-04 model.

Scenario	$\langle\alpha\rangle$	$\mathcal{H}$	Probability	$\ln\mathcal{L}(\langle\alpha\rangle)$
p-He-Li-N-Si-Fe	$(0.51, 0.07, 0.07, 0.11, 0.2, 0.03)$	5.69	0.46	-15378.92
p-N-Fe	$(0.54, 0.39, 0.08)$	4.64	0.45	-15379.93
p-He-N-Fe	$(0.51, 0.05, 0.35, 0.08)$	5.52	0.09	-15380.67
p-N	$(0.49, 0.51)$	2.65	≈ 0	-15393.99
p-He-N	$(0.48, 0.02, 0.5)$	4.97	≈ 0	-15395.01

TABLE 5.9: Same as TABLE 5.7 but using EPOS LHC model.

By looking at TABLE 5.8 we can observe that using QGSJETII-04 model the difference between the probabilities of the best and the second best model is larger than using SIBYLL 2.1 or EPOS LHC. With QGSJETII-04 p-He-N scenario has a probability of 71% and p-He has a probability of 24%. That means that when we analyse with QGSJETII-04 the number of events and the detector resolution are enough to distinguish between three elements. In other words, it is necessary to have three elements to fit the data sample in this energy range. This behaviour is different than when we use SIBYLL 2.1 . In that case the probability of the scenario with two and three elements are almost equal. Besides the likelihood ratio is also larger for QGSJETII-04 getting $\mathcal{L}(\langle\alpha\rangle)_{p-He-N}/\mathcal{L}(\langle\alpha\rangle)_{p-He} = 11.6$. In the case of

QGSJetII-04 the second scenario with largest likelihood is the p-He-N-Fe scenario but the probability of such scenario is negligible against the p-He-N scenario due to the increasing in the composition space. We notice that the estimated composition in the p-He-N and p-He-N-Fe scenarios are almost equal being the iron fraction in the later, *i.e.*, iron is not needed to describe the data and for this reason p-He-N is preferred.

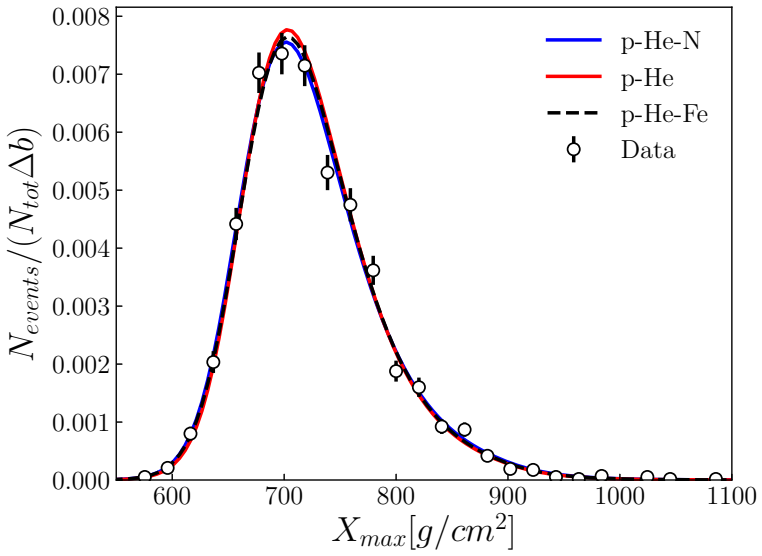


FIGURE 5.38: Same as FIGURE 5.37 but using QGSJetII-04 hadronic interaction model.

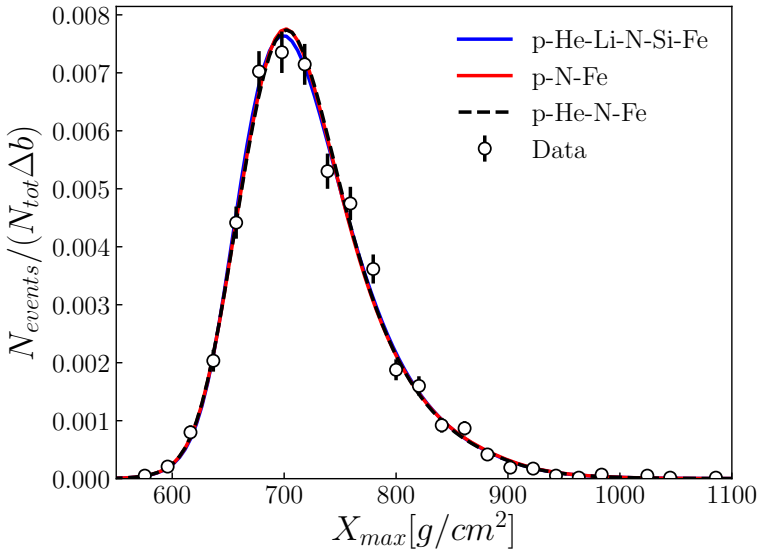


FIGURE 5.39: Same as FIGURE 5.37 but using EPOS LHC hadronic interaction model.

The most surprising result in the model selection is when EPOS LHC is used. In this case the preferred model is the most complex one: the p-He-Li-N-Si-Fe scenario. It is true that the preference for this scenario over the p-N-Fe is not so much larger but remember that the volume of the six-primary scenario is 60 times larger than the three-primary scenario and this increment in the volume is penalised in the Bayesian model selection. In other words, there must be a good reason to explain this result. The easiest explanation for this result is that EPOS LHC really needs the presence of lithium and/or silicon to fit the data. Moreover, this is not the only difference between EPOS LHC and the other models. As was said in SECTION 5.4, while QGSJETII-04 and SIBYLL 2.1 show roughly a transition from lighter to heavier elements, EPOS LHC needs a substantial fraction of nitrogen at all energies. All of these results make us wonder if the combinations of proton, helium, nitrogen and iron used up to now are good enough to describe the data and to compare the models.

To answer this question we first compare the distributions of EPOS LHC with QGSJETII-04 . The mean, standard deviation and mode of the $X_{\max}$ distribution as a function of the energy for these two models are shown in FIGURE 5.40. One can observe that in all energies the following relations between EPOS LHC and QGSJETII-04 are satisfied:

$$\left.\begin{aligned} \langle X_{\max}(A)\rangle_{\text{EPOS LHC}} &\sim \langle X_{\max}(A/2)\rangle_{\text{QGSJETII-04}} \\ \sigma(X_{\max}(A))_{\text{EPOS LHC}} &\sim \sigma(X_{\max}(2A))_{\text{QGSJETII-04}} \\ \text{Mode}(X_{\max}(A))_{\text{EPOS LHC}} &\sim \text{Mode}(X_{\max}(A/2))_{\text{QGSJETII-04}} \end{aligned}\right\} \tag{5.7}$$

We compare the $X_{\max}$ distributions for silicon and lithium generated with EPOS LHC with the distributions of nitrogen and helium generated with QGSJETII-04 in FIGURE 5.41. Moreover, we compare the $X_{\max}$ distributions of nitrogen and iron generated with QGSJETII-04 with a "shifted" $X_{\max}$ distributions of lithium and silicon generated with EPOS LHC in FIGURE 5.42. The "shift" is done in order to fix the mean value of the EPOS LHC $X_{\max}$ distributions to get the same mean value that those generated with QGSJETII-04, *i.e.*, $\langle X_{\max}\rangle^{\text{shifted}}_{\text{EPOS LHC}} \sim 30$ g/cm^2 + $\langle X_{\max}\rangle_{\text{EPOS LHC}}$.

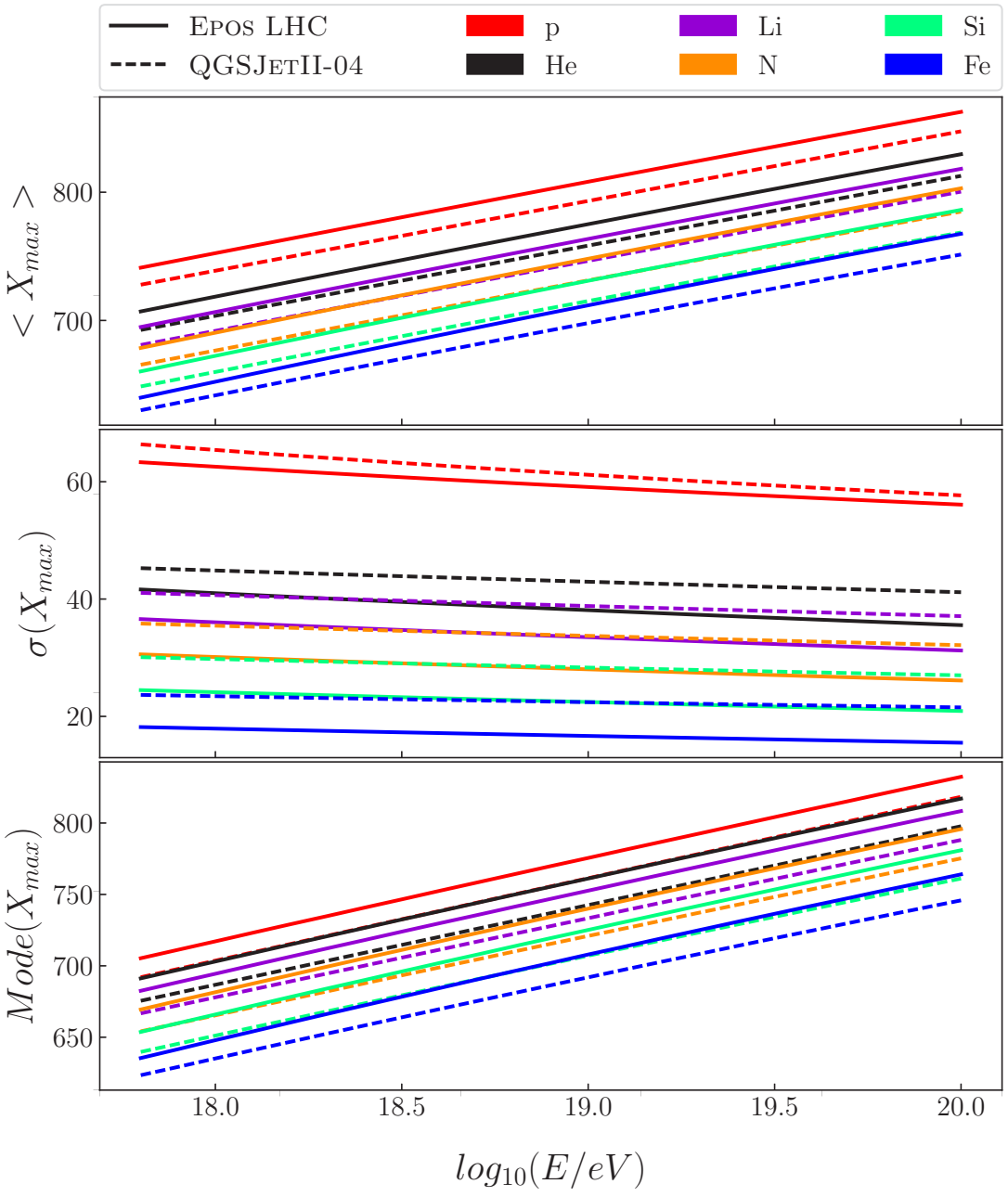


Figure 5.40: Comparison of $X_{\max}$ average, standard deviation and mode for Epos LHC and QGSJetII-04 models for different primaries without detector effects.

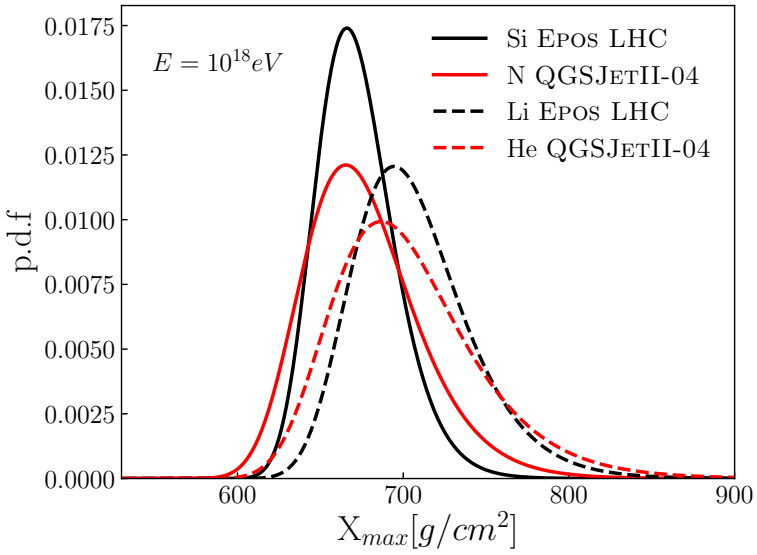


FIGURE 5.41: $X_{\max}$ distributions of silicon and nitrogen generated with EPOS LHC model and nitrogen and helium generated with QGSJETII-04 model without detector effects at 1 EeV.

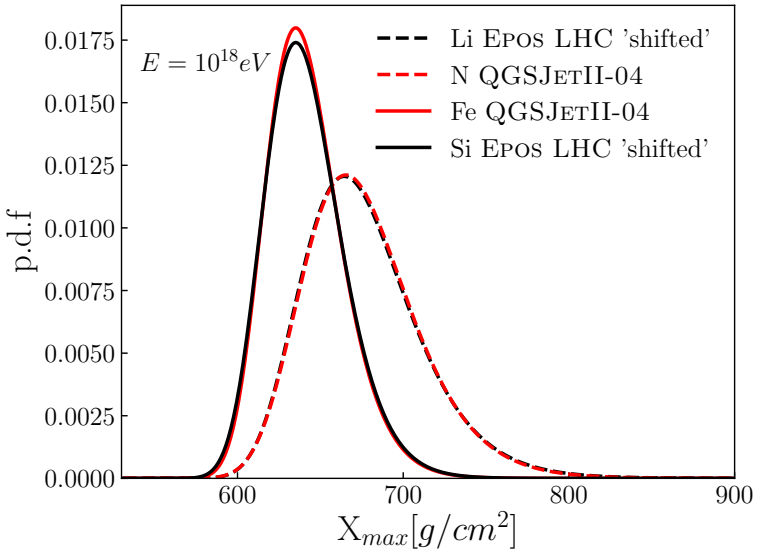


FIGURE 5.42: $X_{\max}$ distributions of silicon and nitrogen generated with EPOS LHC model and nitrogen and helium generated with QGSJETII-04 model without detector effects at 1 EeV.

One can observe that the "shifted" EPOS LHC distributions and the original QGSJETII-04 distributions are almost identical. That could explain why the composition inferred using EPOS LHC is systematically heavier than those inferred by the other models. By comparing FIGURES 5.32-5.33 one can observe that the helium and

lithium fractions are smaller using Epos LHC than using QGSJetII-04 being larger the inferred fractions of nitrogen and silicon are larger.

Since the presence of lithium and silicon has been taken into account only in one primary scenario with six components (which is penalised in the Bayesian model selection) we perform in the next section some "extra scenarios" taking into account more possible combinations with different number of primaries.

5.7 Extra primary scenarios

In this section we analyse the data introducing some extra combination of primaries which were not considered in the previous sections: He-Si, He-N-Si, p-He-Li, p-He-Si, p-Li-N, p-N-Si, p-Li-Si-Fe, N-Si-Fe and Si-Fe. There are two main motivations to explore these new choices: the apparent need of Epos LHC of lithium and silicon to fit the data using Epos LHC in sight of Table 5.6; and the difference in the trends of composition fractions with the energy. These trends seem to be roughly from lighter to heavier elements in Sibyll 2.1 and QGSJetII-04 (with a local maximum of protons around $\log_{10}(E/\text{eV}) = 18.4$) models but while with Epos LHC there is a significant fraction of nitrogen (in the four primary scenario) with a helium fraction which is small in all the energy range and could be neglected.

Naturally, the composition inferred with Epos LHC is expected to be heavier than those given by QGSJetII-04 and Sibyll 2.1 but now we are interested in possible transitions between elements similar to what is observed when QGSJetII-04 and Sibyll 2.1 are used.

By exchanging helium by lithium and nitrogen by silicon the composition fractions change for the three hadronic interaction models as shown in Figures 5.43-5.45. While in the p-He-N-Fe scenario the variation of the proton fraction using Epos LHC is into a variation of the nitrogen fraction in the p-Li-Si-Fe scenario the variation of the proton fraction is translated into a variation of the lithium fraction. Moreover, the iron fraction becomes negligible in this new scenario except for the higher energies. Nevertheless the proton fraction does not change in these two scenarios for any of the hadronic interaction models.

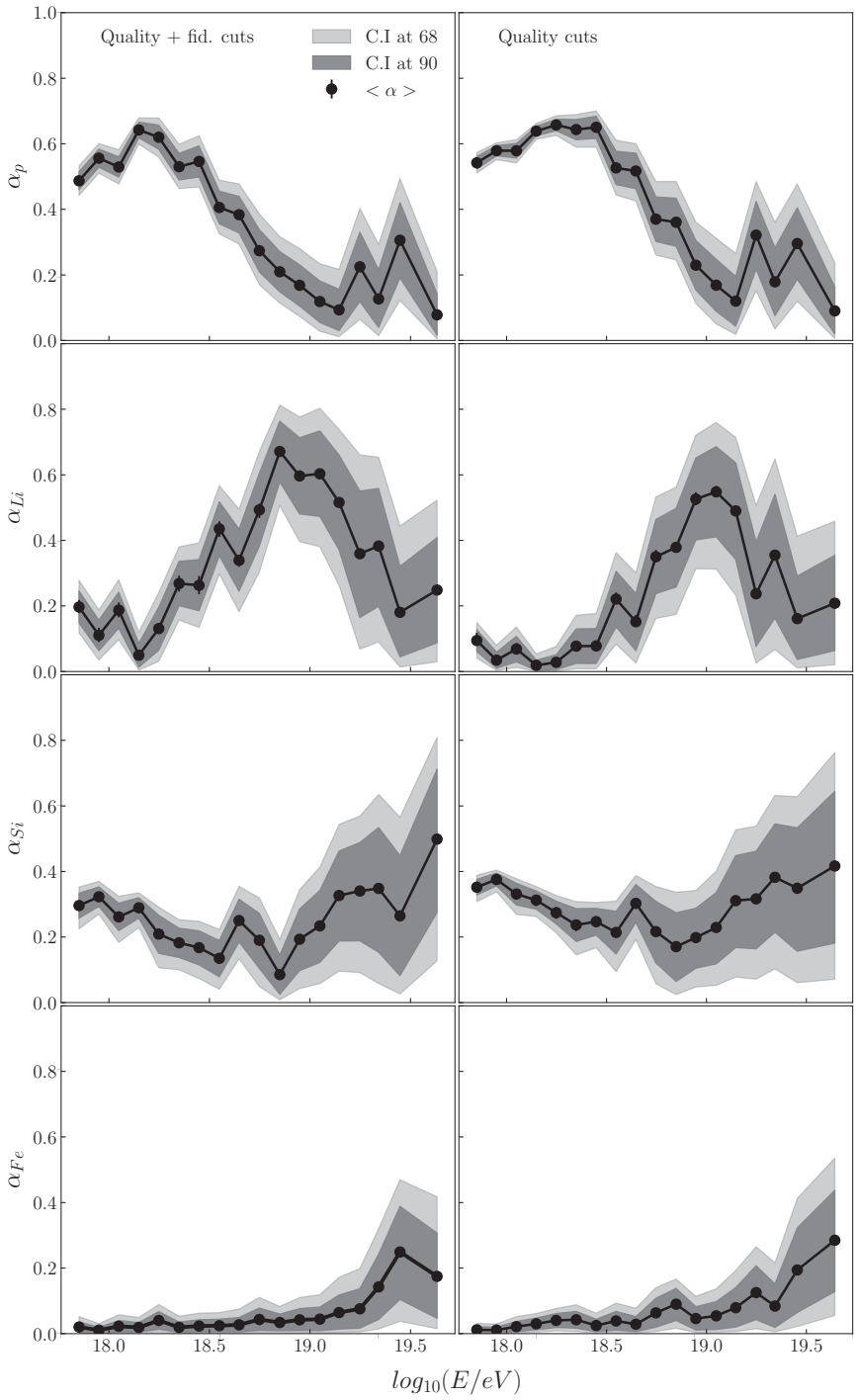


FIGURE 5.43: Proton (first row), helium (second row) nitrogen (third row) and iron (fourth row) trends using EPOS LHC hadronic interaction model. In the left panels the results are obtained using the data sample with fiducial cuts. The estimations using the data sample without fiducial cuts are shown in the right panels.

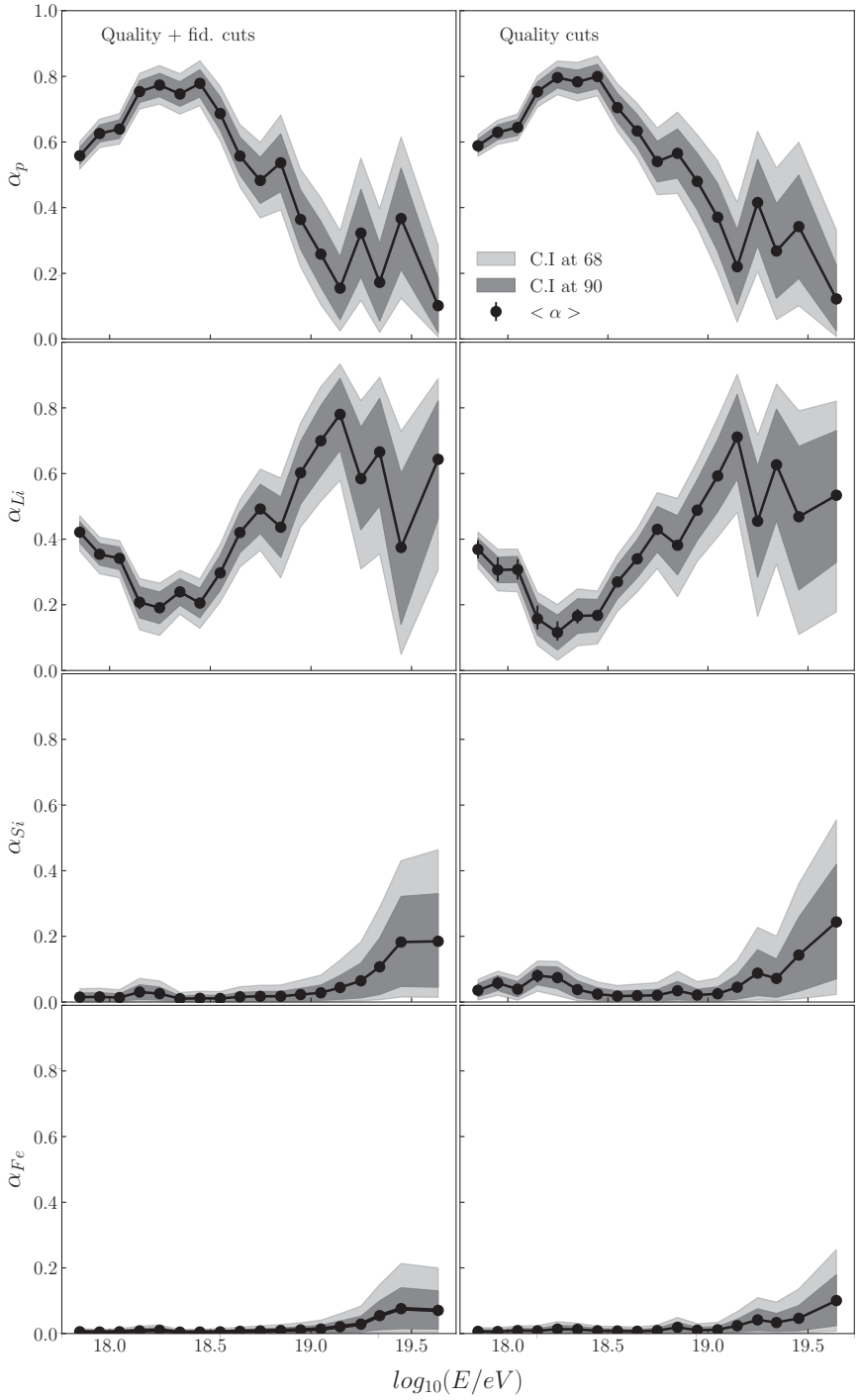


FIGURE 5.44: Same as FIGURE 5.43 but using QGSJETII-04 hadronic interaction model.

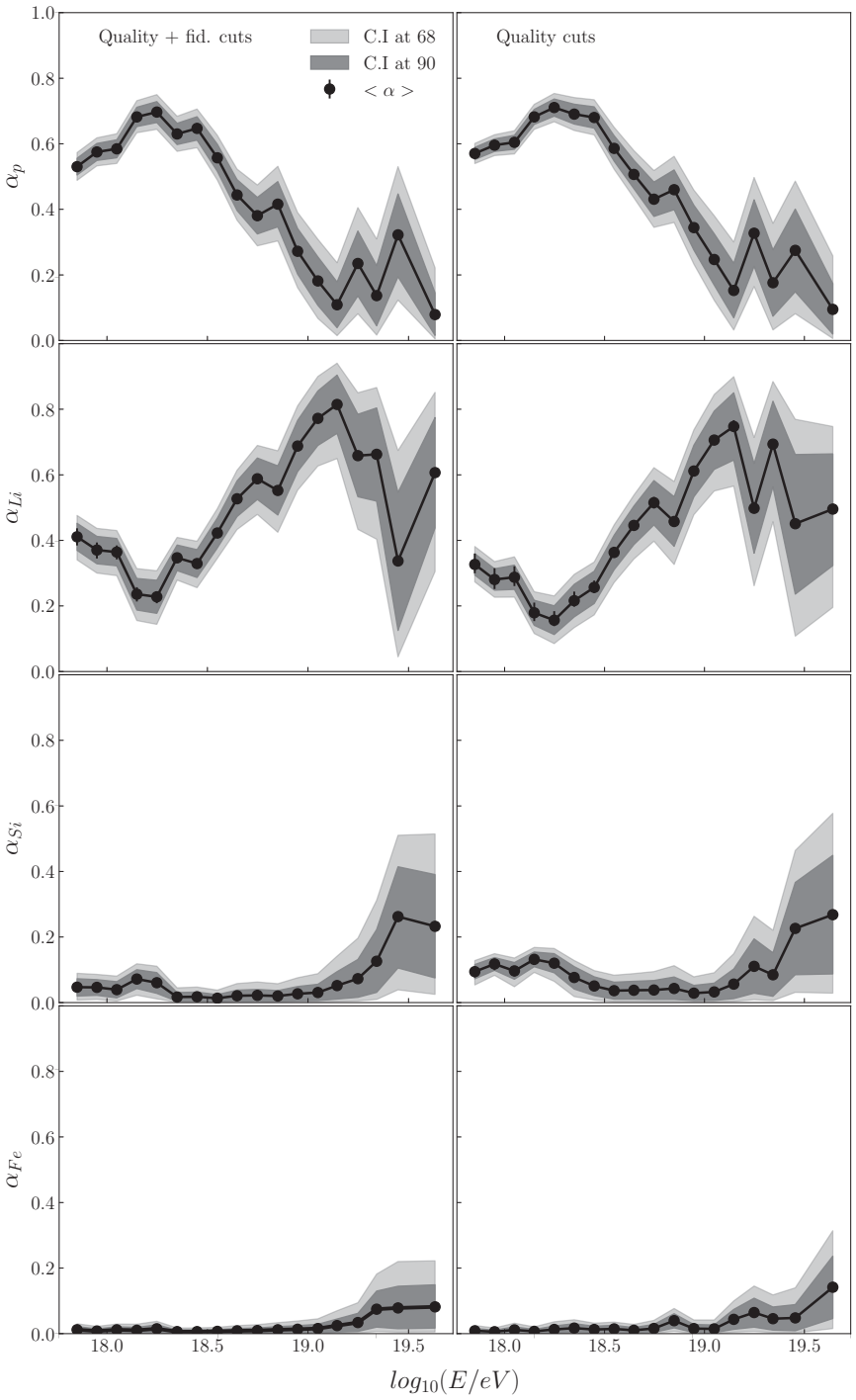


FIGURE 5.45: Same as FIGURE 5.43 but using SIBYLL 2.1 hadronic interaction model.

5.8 Discussion of the scenarios II

As in Section 5.6 we show the best primary scenario for each hadronic model and the best hadronic model for each energy range but taking into account the "extra scenarios" of the previous section in Table 5.10. For each hadronic interaction model there are 21 different primary scenarios. The best three scenarios are shown in Table 5.11.

$\Delta\log_{10}(E/\text{eV})$	Epos LHC	QGSJetII-04	Sibyll 2.1	Best Model
$[17.8, 17.9)$	p-N-Si	p-Li-N	p-Li-N	Epos LHC
$[17.9, 18)$	p-N-Si	p-Li-N	p-Li-N	Epos LHC
$[18, 18.1)$	p-N-Si	p-He-Li	p-Li-N	Epos LHC
$[18.1, 18.2)$	p-N-Si	p-N	p-N	Epos LHC
$[18.2, 18.3)$	p-N-Si	p-N	p-N	QGSJetII-04
$[18.3, 18.4)$	p-N-Si	p-He	p-Li-N	Epos LHC
$[18.4, 18.5)$	p-He-Si	p-He	p-He-Li	QGSJetII-04
$[18.5, 18.6)$	p-N	p-He	p-He-Li	Epos LHC
$[18.6, 18.7)$	p-N-Si	p-He	p-Li-N	Epos LHC
$[18.7, 18.8)$	p-He-Si	p-He	p-He-Li	Sibyll 2.1
$[18.8, 18.9)$	p-Li-N	p-He	p-He	Epos LHC
$[18.9, 19)$	p-N	p-He	p-He-Li	Epos LHC
$[19, 19.1)$	p-N	p-He-Li	p-He-Li	Epos LHC
$[19.1, 19.2)$	p-N	p-He-Li	He-N	Epos LHC
$[19.2, 19.3)$	p-N-Si	p-He-Li	p-Li-N	Epos LHC
$[19.3, 19.4)$	p-He-Li-N-Si-Fe	He-N	He-N	Sibyll 2.1
$[19.4, 19.5)$	p-Li-Si-Fe	p-Li-N	p-N	Epos LHC
$[19.5, \infty)$	N-Si-Fe	He-N	He-N	Epos LHC

Table 5.10: Best primary scenario for each hadronic interaction model and best model for each energy range as in Table 5.6.

$\Delta\log_{10}(E/\text{eV})$	Best scenario (P)	Second (P)	Third (P)
[17.8, 17.9)	EPOS LHC p-N-Si (0.72)	EPOS LHC p-He-Si (0.08)	EPOS LHC p-N-Fe (0.06)
[17.9, 18.0)	EPOS LHC p-N-Si (0.91)	EPOS LHC p-He-Si (0.08)	EPOS LHC p-Li-Si-Fe (0.01)
[18.0, 18.1)	EPOS LHC p-N-Si (0.31)	EPOS LHC p-He-Si (0.24)	QGSJETII-04 p-He-Li (0.22)
[18.1, 18.2)	EPOS LHC p-N-Si (0.52)	EPOS LHC p-He-Si (0.29)	QGSJETII-04 p-N (0.08)
[18.2, 18.3)	QGSJETII-04 p-N (0.27)	QGSJETII-04 p-Li-N (0.23)	QGSJETII-04 p-He-N (0.16)
[18.3, 18.4)	EPOS LHC p-N-Si (0.45)	EPOS LHC p-N (0.25)	EPOS LHC p-N-Fe (0.09)
[18.4, 18.5)	QGSJETII-04 p-He (0.21)	EPOS LHC p-He-Si (0.2)	EPOS LHC p-N-Si (0.14)
[18.5, 18.6)	EPOS LHC p-N (0.4)	EPOS LHC p-LiN (0.16)	EPOS LHC p-He-N (0.14)
[18.6, 18.7)	EPOS LHC p-N-Si (0.43)	EPOS LHC p-N (0.26)	EPOS LHC p-N-Fe (0.07)
[18.7, 18.8)	SIBYLL 2.1 p-He-Li (0.37)	EPOS LHC p-He-Si (0.13)	SIBYLL 2.1 p-He-N (0.1)
[18.8, 18.9)	EPOS LHC p-Li-N (0.35)	EPOS LHC p-He-N (0.18)	SIBYLL 2.1 p-He (0.11)
[18.9, 19.0)	EPOS LHC p-N (0.31)	EPOS LHC p-Li-N (0.22)	EPOS LHC p-He-N (0.15)
[19.0, 19.1)	EPOS LHC p-N (0.3)	EPOS LHC p-Li-N (0.16)	EPOS LHC p-He-N (0.12)
[19.1, 19.2)	EPOS LHC p-N (0.15)	EPOS LHC He-N-Si (0.13)	EPOS LHC p-N-Si (0.12)
[19.2, 19.3)	EPOS LHC p-N-Si (0.12)	SIBYLL 2.1 p-Li-N (0.08)	SIBYLL 2.1 p-He-Li (0.07)
[19.3, 19.4)	SIBYLL 2.1 He-N (0.08)	QGSJETII-04 He-N (0.07)	SIBYLL 2.1 p-Li-N (0.06)
[19.4, 19.5)	EPOS LHC p-Li-Si-Fe (0.1)	EPOS LHC p-N-Fe (0.08)	EPOS LHC p-He-N-Fe (0.07)
[19.5, 21.0)	EPOS LHC N-Si-Fe (0.14)	EPOS LHC N-Fe (0.11)	EPOS LHC He-N-Si (0.1)

TABLE 5.11: Best three scenarios for each energy range with their posterior odds (P).

With the addition of the new primary scenarios we can observe that in the energy range $18 \leq \log_{10}(E/\text{eV}) < 18.1$ the best model scenario for EPOS LHC is not p-He-Li-N-Si-Fe. Now the best scenario is just obtained with p-N-Si. The presence of silicon has reduced the necessity of iron at these energy and a scenario with a less complex space (three primaries instead of six) is preferred. The best five primary scenarios for EPOS LHC are shown in TABLE 5.12 as was done in TABLE 5.9. In sight of the results it is clear that the presence of iron is suppressed by the inclusion of silicon. The posterior odds for p-N-Si and p-He-Si are much larger than the other scenarios. Notice that the six-component scenario is not only penalised due to its complexity (dimension) but also because the likelihoods of p-N-Si and p-He-Si are larger.

Scenario	$\langle\alpha\rangle$	$\mathcal{H}$	Probability	$\ln\mathcal{L}(\langle\alpha\rangle)$
p-N-Si	$(0.57, 0.19, 0.24)$	3.54	0.5	-15378.25
p-He-Si	$(0.51, 0.17, 0.33)$	3.66	0.39	-15378.37
p-Li-Si-Fe	$(0.53, 0.19, 0.26, 0.02)$	5.47	0.05	-15378.5
p-He-Li-N-Si-Fe	$(0.51, 0.07, 0.07, 0.11, 0.2, 0.03)$	5.69	0.03	-15378.92
p-N-Fe	$(0.54, 0.39, 0.08)$	4.64	0.03	-15379.93

TABLE 5.12: The five best primary scenarios considering the new cases using EPOS LHC. The energy range is $18 \leq \log_{10}(E/\text{eV}) < 18.1$.

Now we explore the last three. In these bins it seems that the proton fraction disappears in the energy bin $\log_{10}(E/\text{eV}) \sim 19.3$, appears at $\log_{10}(E/\text{eV}) \sim 19.4$ and disappears again at the highest energies. In bin $19.3 \leq \log_{10}(E/\text{eV}) < 19.4$ the best scenario is SIBYLL 2.1 He-N. If we compare the posterior odds of the best scenario with the second and third best scenarios we see that the preference for this scenario over QGSJETII-04 He-N and SIBYLL 2.1 p-Li-N is marginal. In the other two energy bins a similar situation takes place, with the difference that EPOS LHC is now the preferred model.

We inspect the different scenarios for the three hadronic interaction models separately in TABLES 5.13-5.15. We can observe that the presence of proton in the energy range $19.4 \leq \log_{10}(E/\text{eV}) < 19.5$ is necessary to fit the data for all the hadronic models with fractions between 20% and 50% with QGSJETII-04 and SIBYLL 2.1; and between 50% and 60% for *epos*. However, in the two adjacent bins protons seem not to be necessary. In the few cases where an scenario contains protons its fraction is less than 20%. Moreover, in the cases where one of the scenarios appearing in the tables contains protons, its fraction is less than 20%. One of the possible explanations for this behaviour could be a source of protons emitting at energies around $19.4 \leq \log_{10}(E/\text{eV})19.5$ but this possible explanation should be produce an anisotropy in the direction of the sources due to the energy of the particles. We explore possible proton flux in CHAPTER 6. Of course, a statistical fluctuation can be also the explanation of this behaviour.

Scenario	$\langle\alpha\rangle$	$\mathcal{H}$	Probability	$\ln\mathcal{L}(\langle\alpha\rangle)$
$[19.3 \leq \log_{10}(E/\text{eV}) < 19.4)$				
p-He-Li-N-Si-Fe	$(0.09, 0.14, 0.17, 0.23, 0.22, 0.14)$	0.71	0.15	-321.41
He-N-Si	$(0.29, 0.31, 0.41)$	0.56	0.14	-321.68
He-Si	$(0.41, 0.59)$	0.83	0.14	-321.54
p-N-Si	$(0.17, 0.43, 0.4)$	0.76	0.12	-321.56
He-N-Fe	$(0.3, 0.49, 0.21)$	0.85	0.11	-321.44
$[19.4 \leq \log_{10}(E/\text{eV}) < 19.5)$				
p-Li-Si-Fe	$(0.31, 0.18, 0.26, 0.25)$	0.75	0.18	-206.38
p-Fe	$(0.36, 0.29, 0.36)$	0.88	0.14	-206.42
p-He-N-Fe	$(0.25, 0.19, 0.21, 0.35)$	0.85	0.12	-206.68
p-He-Li-N-Si-Fe	$(0.21, 0.13, 0.12, 0.13, 0.18, 0.23)$	0.75	0.12	-206.8
p-He-Fe	$(0.28, 0.29, 0.44)$	0.88	0.11	-206.85
$[19.5 \leq \log_{10}(E/\text{eV}) < \infty)$				
N-Si-Fe	$(0.5, 0.34, 0.16)$	0.55	0.22	-177.2
N-Fe	$(0.74, 0.26)$	0.68	0.17	-177.31
He-N-Si	$(0.13, 0.32, 0.55)$	0.74	0.15	-177.41
He-Si	$(0.2, 0.8)$	0.8	0.14	-177.42
p-N-Si	$(0.08, 0.36, 0.56)$	1.11	0.08	-177.65

TABLE 5.13: From left to right: best primary scenario, inferred composition, gain of information with respect to the prior, posterior odds and log-likelihood of the inferred composition using EPOS LHC for the energies: $19.3 \leq \log_{10}(E/\text{eV}) < 19.4$, $19.4 \leq \log_{10}(E/\text{eV}) < 19.5$ and $19.5 \leq \log_{10}(E/\text{eV}) < \infty$

Scenario	$\langle\alpha\rangle$	$\mathcal{H}$	Probability	$\ln\mathcal{L}(\langle\alpha\rangle)$
He-N	$(0.65, 0.35)$	0.32	0.24	-321.46
p-Li-N	$(0.17, 0.53, 0.3)$	0.57	0.14	-321.72
p-He-Li	$(0.1, 0.31, 0.59)$	0.9	0.12	-321.74
p-He-N	$(0.14, 0.44, 0.42)$	0.65	0.1	-321.92
He-Si	$(0.82, 0.18)$	0.89	0.09	-321.92
$[19.4 \leq \log_{10}(E/\text{eV}) < 19.5)$				
p-Li-N	$(0.33, 0.35, 0.33)$	0.35	0.15	-208.34
p-He-N	$(0.28, 0.34, 0.38)$	0.32	0.14	-208.46
p-N	$(0.43, 0.57)$	0.61	0.12	-208.27
p-He-Li	$(0.22, 0.34, 0.44)$	0.27	0.11	-208.93
p-He-Si	$(0.3, 0.43, 0.27)$	0.39	0.08	-208.92
$[19.5 \leq \log_{10}(E/\text{eV}) < \infty)$				
He-N	$(0.36, 0.64)$	0.18	0.31	-179.72
p-Li-N	$(0.09, 0.44, 0.47)$	0.8	0.14	-179.94
He-N-Si	$(0.37, 0.46, 0.17)$	0.38	0.13	-180.33
p-N	$(0.13, 0.87)$	1.08	0.07	-180.38
p-He-N	$(0.1, 0.29, 0.62)$	1.03	0.07	-180.4

TABLE 5.14: Same as TABLE 5.13 but using QGSJETII-04 hadronic interaction model.

Scenario	$\langle\alpha\rangle$	$\mathcal{H}$	Probability	$\ln\mathcal{L}(\langle\alpha\rangle)$
$[19.3 \leq \log_{10}(E/\text{eV}) < 19.4)$				
He-N	$(0.5, 0.5)$	0.45	0.19	-321.19
p-Li-N	$(0.15, 0.45, 0.4)$	0.67	0.14	-321.26
He-N-Si	$(0.56, 0.28, 0.16)$	0.91	0.12	-321.2
He-Si	$(0.71, 0.29)$	0.88	0.1	-321.43
p-He-N	$(0.13, 0.32, 0.55)$	1.0	0.09	-321.39
$[19.4 \leq \log_{10}(E/\text{eV}) < 19.5)$				
p-N	$(0.35, 0.65)$	0.82	0.24	-207.0
p-N-Si	$(0.38, 0.4, 0.22)$	0.75	0.17	-207.35
p-Li-N	$(0.29, 0.26, 0.46)$	0.64	0.17	-207.5
p-He-N	$(0.26, 0.2, 0.54)$	0.86	0.12	-207.54
p-He-Si	$(0.29, 0.3, 0.41)$	0.66	0.07	-208.24
$[19.5 \leq \log_{10}(E/\text{eV}) < \infty)$				
He-N	$(0.24, 0.76)$	0.53	0.31	-177.77
He-N-Si	$(0.28, 0.53, 0.19)$	0.48	0.16	-178.33
p-Li-N	$(0.08, 0.34, 0.58)$	1.11	0.12	-178.22
p-N	$(0.1, 0.9)$	1.31	0.09	-178.23
He-N-Fe	$(0.27, 0.65, 0.08)$	1.29	0.06	-178.56

TABLE 5.15: Same as TABLE 5.13 but using SIBYLL 2.1 hadronic interaction model.

5.9 Comments on the hadronic interaction models

Looking at TABLE 5.6 and TABLE 5.11 the conclusion about which of the hadronic models represents better the data is clear: EPOS LHC is the best hadronic model. Nevertheless, in APPENDIX E we show a study about the evidence in the model selection for simulations with different number of events. In SECTION E.1 we simulate data sets performed with 100, 500, 1000, 2000, 3000 and 5000 events for each of the three hadronic interaction models and the signal is "smeared" using the response function that it is applied for data. The composition in the mock data sets is always the same: $\alpha_p = 1/3$, $\alpha_{He} = 5/18$, $\alpha_N = 2/9$ and $\alpha_{Fe} = 1/6$. These numbers are totally arbitrary but are chosen in this way to be non-zero and to ensure dominance of

lighter elements over heavier. Once the events have been simulated, they are analysed using the same procedure and assuming the same primary scenarios as for the data.

The results are shown in Tables E.1-E.3. One can observe that the best model-scenario coincides with the simulated only when the number of events is larger than 3000. Notice that in the data with fiducial cuts this is only true for the first bin. Below this number of events, the fluctuations affect the selection of the preferred model which appears as an almost random selection. Actually, the posterior odds in the case of simulations does not discriminate so well as in data. For some unknown reason the power to identify the best model-scenario seems to be better in the case of data than in simulations. It is difficult to interpret these results. Is is possible ...

To investigate these results deeper we perform a second set of simulations (Section E.2) but in this case we do not "smear" the mock data sets and again we analyse them just as it is done for the data (assuming a response function). The mock data set is built with distributions whose widths are artificially smaller than those taken into account at the time of the analysis. The results are shown in Tables E.4-E.5 where one can see that Epos LHC is the preferred model by all simulations independently of the number of events in the data set or of the hadronic model used to generate them.

Summarising we can conclude that if the detector is well understood and one of the models describes well the high energy interactions, we would not see the preference for Epos LHC such as that observed in data. Our detector is not able to distinguish the true hadronic model with the actual number of events available in data. Nevertheless, if the true $X_{\max}$ distributions are narrower than what we believe Epos LHC is preferred (as it is observed when we analyse actual data). It can be well understood in sight of Figure 5.40. Since Epos LHC $X_{\max}$ distributions have the smallest standard deviations it has more "freedom" to fit the data than QGSJetII-04 and Sibyll 2.1. That could also be the reason why Epos LHC can reproduce better the standard deviations of data in the posterior predictive distributions while QGSJetII-04 and Sibyll 2.1 show larger values. If we trust in the response function used in the official analysis (and we should) we must conclude that none of the models describe the high energy interactions that are producing the air showers.

5.10 The proton fraction: robust behaviour

To extract conclusions about the composition comparing all primary scenarios and treating the hadronic models separately we compare the proton fractions obtained in the six-component scenario with the p-He-N-Fe scenario and with the best scenario for each energy bin. This is plotted as a function of the three hadronic models in FIGURE 5.46. We choose p-He-Li-N-Si-Fe and p-He-N-Fe because they are the most complete, the former from this work and the latter from published work by the Pierre Auger Collaboration. One can observe that the proton fraction can be inferred quite well because the three scenarios give us compatible results within each hadronic model (note that the best scenario changes from one bin to another). We also notice that the six and four-primary scenarios could be no compatible with the best scenario in some bins, but taking a look at the posterior probability we observe that they are compatible. For instance, we show in FIGURE 5.47 the posterior probability density function of the proton fraction in the energy bin $19.3 \leq \log_{10}(E/\text{eV}) < 19.4$ for QGSJETII-04 and SIBYLL 2.1 where the best scenario says that no protons are needed.

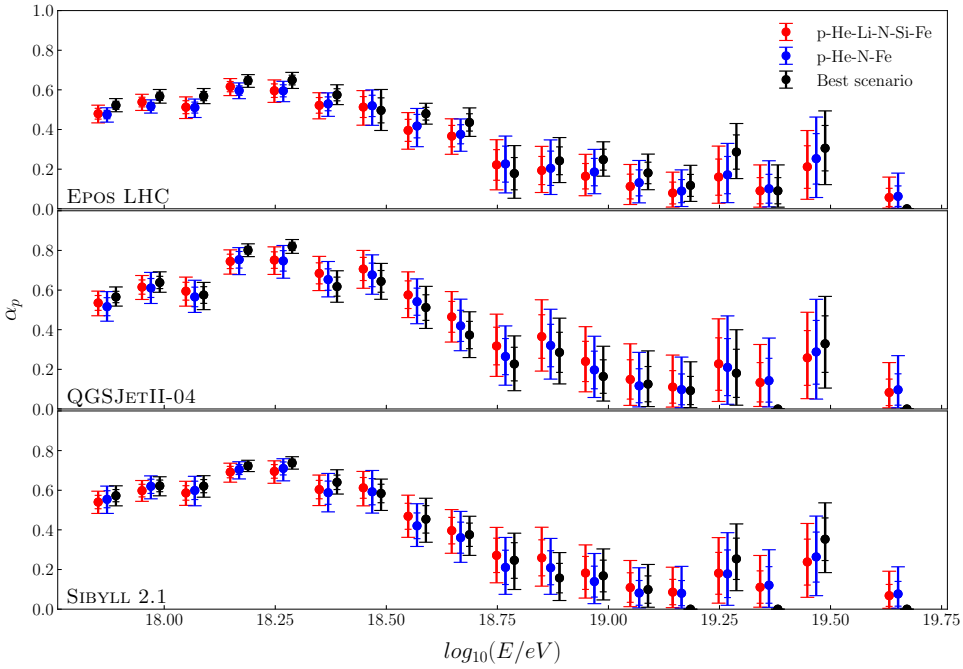


FIGURE 5.46: Comparison of the proton fraction obtained with the p-He-Li-N-Si-Fe (red), p-He-N-Fe (blue) and the best (black) primary scenarios for the three hadronic interaction models: EPOS LHC (upper panel), QGSJETII-04 (middle) and SIBYLL 2.1 (lower).

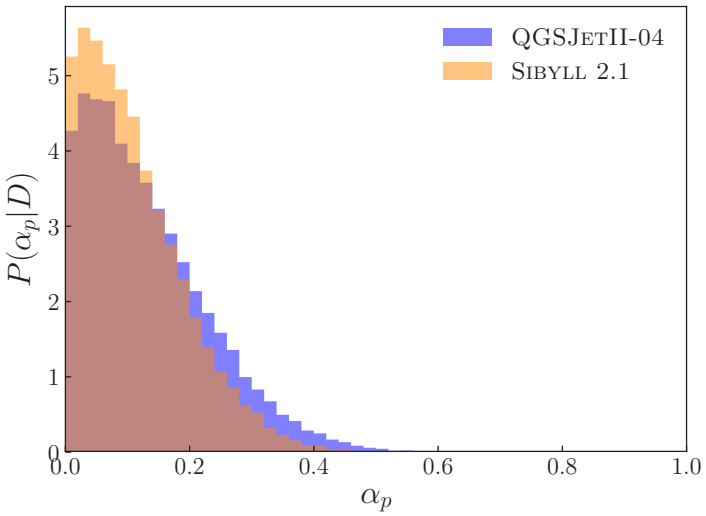


FIGURE 5.47: Posterior proton probability density function for QGSJETII-04 (blue) and SIBYLL 2.1 (orange) in the energy bin $19.3 \leq \log_{10}(E/\text{eV}) < 19.4$.

We observe that the inferred proton fraction has a robust behaviour for a give hadronic model. As all possible primary scenarios are subsets of the six-primary scenario, we present our final result on the proton fraction with this scenario. We finally compare the proton fractions obtained with the three hadronic models in FIGURE 5.48.

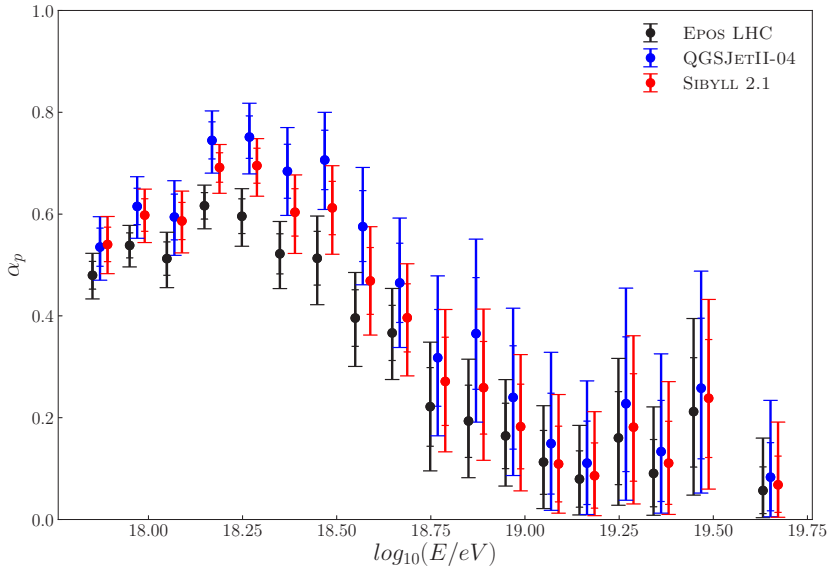


FIGURE 5.48: Average of the posterior probability density functions of the proton fraction as a function of the energy for EPOS LHC (black), QGSJETII-04 (blue) and SIBYLL 2.1 (red). 65% and 90% confidence intervals of the posterior distribution are respectively denoted by smaller and bigger error bars.

We observe that in all energy bins EPOS LHC gives the smallest proton fraction and the biggest is given by QGSJETII-04. The SIBYLL 2.1 hadronic interaction model is always in the middle. The three hadronic interaction models describe a similar behaviour: at the lowest energy bins in this analysis the proton fraction increases reaching a maximum around the energy range $\log_{10}(E/\text{eV}) \in [18.2, 18.4]$. Once this maximum is reached the proton fraction falls. Beyond 10 EeV the uncertainties are too large to conclude if the protons disappear. None of the hadronic models give us pure proton in any energy bin. Besides, the proton fractions given by the three hadronic models are compatible taking into account the uncertainties.

Another measurement that does not depend on the primary scenario is the logarithm of the mass number as it is shown in FIGURE 5.49 where we compare the results for ln$A$ obtained with the best scenario in each bin, with the p-He-N-Fe scenario and with the p-He-Li-N-Si-Fe scenario . The inferred values of the ln$A$ are even more robust than the inference on the proton fraction.

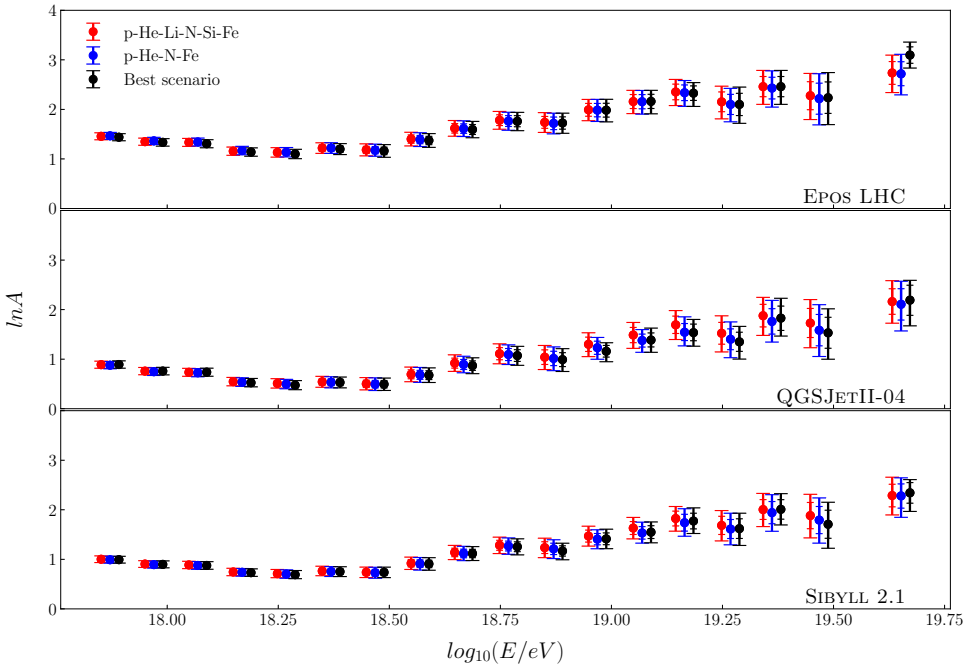


FIGURE 5.49: Same as FIGURE 5.46 but for the posterior distribution of the logarithm of the mass number.

The analysis for the three hadronic models present a minimum in ln$A$ in the same energy region where the proton fraction has its maximum. EPOS LHC gives the

heaviest composition while QGSJetII-04 gives the lightest. Nevertheless, in this case the inference for the three hadronic models are not clearly not compatible as illustrated in Figure 5.50. Sibyll 2.1 and QGSJetII-04 are compatible at energies beyond $\log_{10}(E/\text{eV}) \geq 18.7$ and Epos LHC is compatible with QGSJetII-04 and Sibyll 2.1 beyond $\log_{10}(E/\text{eV}) = 19.4$ where the uncertainties are also larger. The posterior predictive distributions for the logarithm of the masses are shown in Appendix F for each energy bin.

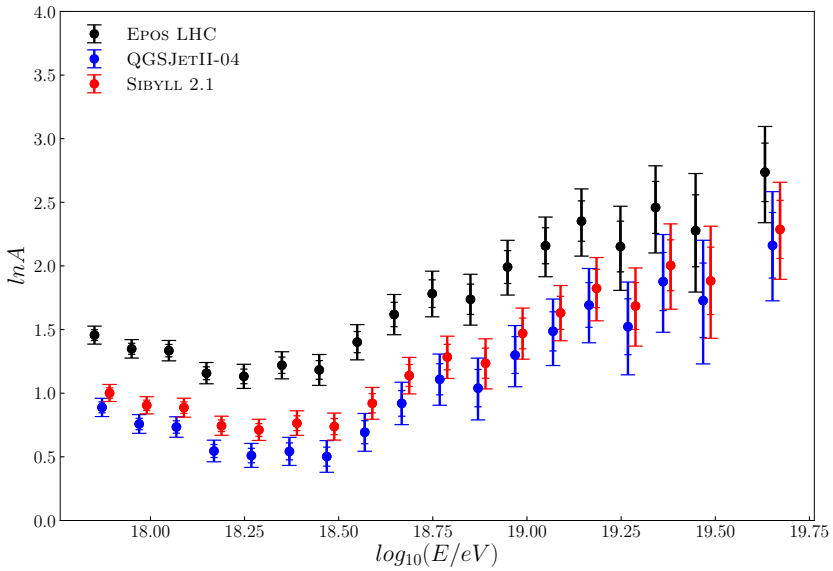


Figure 5.50: Average of the posterior probability distribution of the logarithm of the mass number as a function of the energy for Epos LHC (black), QGSJetII-04 (blue) and Sibyll 2.1 (red). The smaller (larger) error bars denote the 65% (90%) of the integral of the distribution.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 6

Preliminary proton flux

In this chapter we present the preliminary proton and non-proton fluxes measured by the Pierre Auger Observatory. To infer the proton flux we use the parameterisation of the all-particle energy spectrum presented in [42] and the proton fractions obtained in this work using the six-primary scenario (Section 5.5). We find that the proton fluxes inferred by the three hadronic interaction models are compatible.

6.1 Approximations to the proton and non-proton fluxes

The Pierre Auger Combined flux is well described as a power-law below a certain energy (the ankle, denoted by E_a) and a power-law with a smooth suppression at the highest energies:

$$J(E) = J_0 \begin{cases} \left(\frac{E}{E_a}\right)^{-\gamma_1} & \text{if } E \leq E_a \\ \left(\frac{E}{E_a}\right)^{-\gamma_2} \frac{1+\left(\frac{E_a}{E_s}\right)^{\Delta\gamma}}{1+\left(\frac{E}{E_s}\right)^{\Delta\gamma}} & \text{if } E > E_a \end{cases} . \tag{6.1}$$

The value of the best-fit parameters are listed in Table 6.1 with the respective statistical and systematic uncertainties. In this work we only use the best-fit values and the uncertainties are used only for comparison purposes.

$J_0[\mathrm{eV}^{-1}\mathrm{km}^{-2}\mathrm{sr}^{-1}\mathrm{yr}^{-1}]$	E_a[EeV]	E_s[EeV]	γ_1	γ_2	$\Delta\gamma$
$(3.30 \pm 0.15 \pm 0.20) \times 10^{-19}$	$4.82 \pm 0.07 \pm 0.8$	$42.09 \pm 1.7 \pm 7.61$	$3.29 \pm 0.002 \pm 0.05$	$2.60 \pm 0.02 \pm 0.1$	$3.14 \pm 0.2 \pm 0.4$

TABLE 6.1: Best-fit parameters, with statistical and systematic uncertainties, for the combined energy spectrum measured at the Pierre Auger Observatory. Taken from [42].

To infer the proton flux we use the marginal proton posterior probability density functions obtained in the p-He-Li-N-Si-Fe scenario for each hadronic model by approximating these distributions by normal distributions with mean and standard deviations equal to those of the marginals. The non-proton flux is then obtained by subtracting the proton flux from the total flux. All fluxes are shown in FIGURE 6.1.

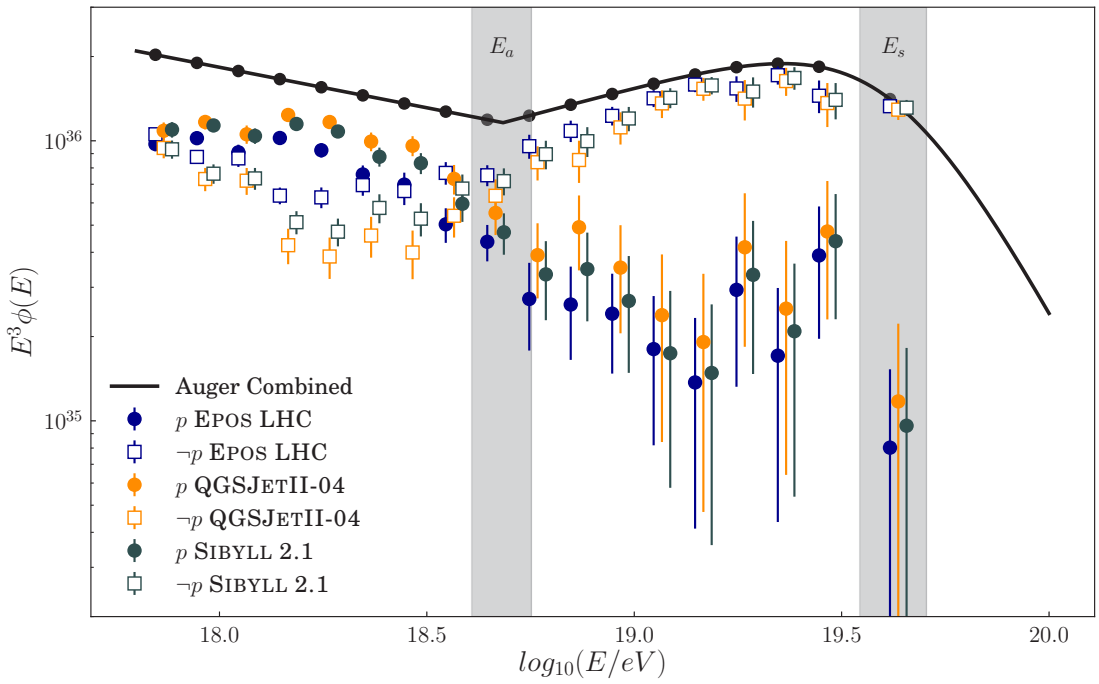


FIGURE 6.1: The Auger combined spectrum (solid line) and proton (non-proton) spectra with circles (squares) for EPOS LHC (blue), QGSJETII-04 (orange) and SIBYLL 2.1 (green) models. The combined spectrum is normalised to the unity while the proton and non-proton spectra are normalised to the respective fractions.

6.2 Spectral features in the proton and non-proton spectra

The all-particle spectrum above $\log_{10}(E/\text{eV}) = 17.8$ can be describe as:

$$\phi(E) = \eta_p \phi_p(E) + (1 - \eta_p)\phi_{\neg p}(E), \tag{6.2}$$

where η_p is the total proton fraction from $\log_{10}(E/\text{eV}) = 17.8$ up to ∞ using the six-component scenario of the CHAPTER 5. The values of the total proton fraction for each hadronic model are shown in TABLE 6.2. The functions $\phi_p(E)$ and $\phi_{\neg p}(E)$ are the proton and non-proton fluxes respectively normalised to one.

Model	η_p	C.I at 68%	C.I at 90%
EPOS LHC	0.54	$[0.52, 0.55]$	$[0.52, 0.55]$
QGSJETII-04	0.59	$[0.55, 0.61]$	$[0.55, 0.62]$
SIBYLL 2.1	0.58	$[0.57, 0.59]$	$[0.56, 0.6]$

TABLE 6.2: Total proton fraction and its 68% and 90% confidence intervals for each hadronic model.

The characteristic features of the spectra are identified and studied by fitting different models to the inferred fluxes. Each spectrum is normalised to one by removing a degree of freedom due to the normalisation constant because at this point we are interested in the spectral features more than in the total proton flux.

We investigate several functional forms to fit both proton and non-proton spectra, all of which are based on power laws. These are given by EQUATION 6.3 to EQUATION 6.11, and sorted by the number of free fit parameters:

$$\phi_1(E;\gamma) \propto E^{-\gamma}, \tag{6.3}$$

$$\phi_2(E;\gamma, E_c) \propto E^{-\gamma} \exp\left(-E/E_c\right), \tag{6.4}$$

$$\phi_3(E;\gamma_1, \gamma_2, E_s) \propto \frac{E^{-\gamma_1}}{1 + \left(\frac{E}{E_s}\right)^{\gamma_2}}, \tag{6.5}$$

$$\phi_4(E;\gamma_1, \gamma_2, E_b) \propto \begin{cases} E^{-\gamma_1} & \text{if } E \leq E_b \\ K E^{-\gamma_2} & \text{if } E_b < E \end{cases}, \tag{6.6}$$

$$\phi_5(E;\gamma_1,\gamma_2,E_b,E_c) \propto \begin{cases} E^{-\gamma_1} & \text{if } E \leq E_b \\ KE^{-\gamma_2}\exp\left(-E/E_c\right) & \text{if } E_b < E \end{cases}, \tag{6.7}$$

$$\phi_6(E;\gamma_1,\gamma_2,\gamma_3,E_b,E_s) \propto \begin{cases} E^{-\gamma_1} & \text{if } E \leq E_b \\ K\frac{E^{-\gamma_2}}{1+\left(\frac{E}{E_s}\right)^{\gamma_3}} & \text{if } E_b < E \end{cases}, \tag{6.8}$$

$$\phi_7(E;\gamma_1,\gamma_2,E_b,E_b'E_c) \propto \begin{cases} E^{-\gamma_1} & \text{if } E \leq E_b \\ KE^{-\gamma_2} & \text{if } E_b < E \leq E_b' \\ K'E^{-\gamma_3} & \text{if } E_b' < E \end{cases}, \tag{6.9}$$

$$\phi_8(E;\gamma_1,\gamma_2,\gamma_3,E_b,E_b',E_c) \propto \begin{cases} E^{-\gamma_1} & \text{if } E \leq E_b \\ KE^{-\gamma_2} & \text{if } E_b < E \leq E_b' \\ K'E^{-\gamma_3}\exp\left(-E/E_c\right) & \text{if } E_b' < E \end{cases}. \tag{6.10}$$

$$\phi_9(E;\gamma_1,\gamma_2,\gamma_3,\gamma_4,E_b,E_b',E_s) \propto \begin{cases} E^{-\gamma_1} & \text{if } E \leq E_b \\ KE^{-\gamma_2} & \text{if } E_b < E \leq E_b' \\ K'\frac{E^{-\gamma_3}}{1+\left(\frac{E}{E_s}\right)^{\gamma_4}} & \text{if } E_b' < E \end{cases}. \tag{6.11}$$

In Equations 6.12-6.11 K and K' are constants that make the functions continuous at the energy breaks. All the functions are combinations of power-laws with breaks and a cut-off following a power-law or an exponential.

In sight of Figure 6.1 one might expect that the simple power-law function (ϕ_1) will not perform a good fit to the spectra. Nevertheless,we can compare the goodness of the fits for the different functions with the posterior odds.

The posterior odds of the fits normalised to the unity (*i.e.*, the probability) are shown in Tables 6.3-6.5 for the three hadronic models.

Function	Probability
Proton flux	
$\phi_4(E;\gamma_1,\gamma_2,E_b)$	0.22
$\phi_3(E;\gamma_1,\gamma_2,E_s)$	0.16
$\phi_7(E;\gamma_1,\gamma_2,E_b,E_b',E_c)$	0.15
$\phi_8(E;\gamma_1,\gamma_2,\gamma_3,E_b,E_b',E_c)$	0.14
$\phi_5(E;\gamma_1,\gamma_2,E_b,E_c)$	0.12
$\phi_9(E;\gamma_1,\gamma_2,\gamma_3,\gamma_4,E_b,E_b',E_s)$	0.11
Non-proton flux	
$\phi_8(E;\gamma_1,\gamma_2,\gamma_3,E_b,E_b',E_c)$	0.63
$\phi_5(E;\gamma_1,\gamma_2,E_b,E_c)$	0.15

TABLE 6.3: Probability of the fit of the different functions to the proton and non-proton EPOS LHC spectra. Only functions with probability larger than 10% are shown.

Function	Probability
Proton flux	
$\phi_4(E;\gamma_1,\gamma_2,E_b)$	0.21
$\phi_3(E;\gamma_1,\gamma_2,E_s)$	0.2
$\phi_7(E;\gamma_1,\gamma_2,E_b,E_b',E_c)$	0.15
$\phi_8(E;\gamma_1,\gamma_2,\gamma_3,E_b,E_b',E_c)$	0.14
$\phi_5(E;\gamma_1,\gamma_2,E_b,E_c)$	0.12
$\phi_9(E;\gamma_1,\gamma_2,\gamma_3,\gamma_4,E_b,E_b',E_s)$	0.1
Non-proton flux	
$\phi_8(E;\gamma_1,\gamma_2,\gamma_3,E_b,E_b',E_c)$	0.41
$\phi_5(E;\gamma_1,\gamma_2,E_b,E_c)$	0.38

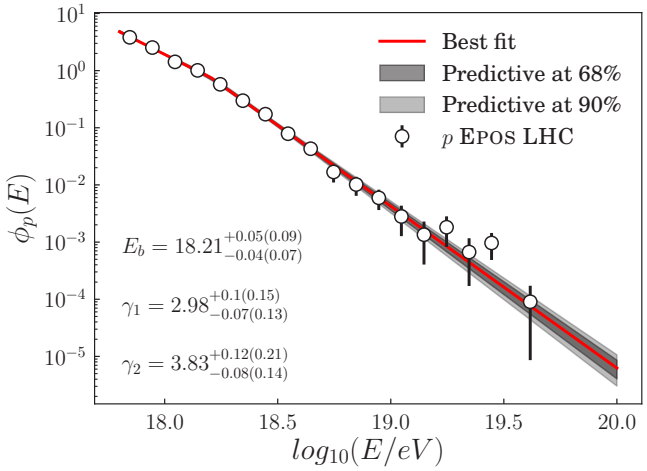
TABLE 6.4: Same as TABLE 6.3 but using QGSJETII-04.

Function	Probability
Proton flux	
$\phi_4(E;\gamma_1,\gamma_2,E_b)$	0.22
$\phi_7(E;\gamma_1,\gamma_2,E_b,E'_b,E_c)$	0.15
$\phi_3(E;\gamma_1,\gamma_2,E_s)$	0.15
$\phi_8(E;\gamma_1,\gamma_2,\gamma_3,E_b,E'_b,E_c)$	0.15
$\phi_5(E;\gamma_1,\gamma_2,E_b,E_c)$	0.12
$\phi_9(E;\gamma_1,\gamma_2,\gamma_3,\gamma_4,E_b,E'_b,E_s)$	0.11
Non-proton flux	
$\phi_8(E;\gamma_1,\gamma_2,\gamma_3,E_b,E'_b,E_c)$	0.4
$\phi_5(E;\gamma_1,\gamma_2,E_b,E_c)$	0.35
$\phi_7(E;\gamma_1,\gamma_2,E_b,E'_bE_c)$	0.11

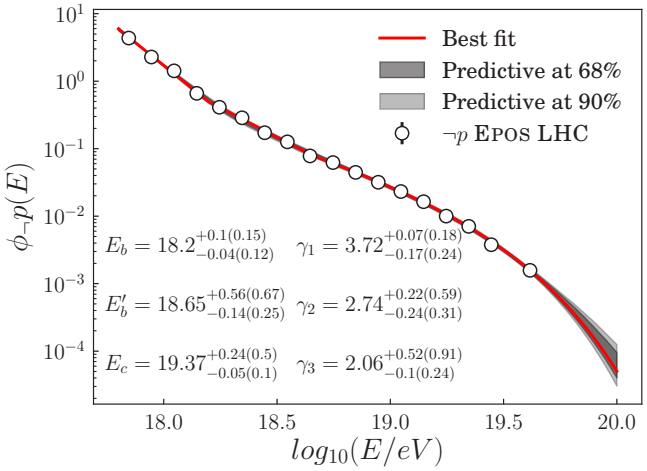
TABLE 6.5: Same as TABLE 6.3 but using SIBYLL 2.1.

The best function to fit the proton spectrum is $\phi_4(E;\gamma_1,\gamma_2,E_b)$, which describes a broken power law function with power $-\gamma_1$ that changes to $-\gamma_2$ at the energy break E_b. The difference in probability between the probabilities for this function and for the second best function depends on the assumed hadronic model. For QGSJETII-04 this difference is small but ϕ_4 is still the most probable function. We notice that when accounting for the uncertainties in the combined spectrum and without approximations in the posterior proton distributions the results could change, but this requires to estimate the combined spectrum itself under a Bayesian approach which is out of the scope of these work. For the moment we assume the fact that for the three hadronic models the best fit is described by the same function. The same reasoning can be applied for the non-proton spectrum. The latter is best described by a broken power law with two breaks and an exponential cut-off.

The proton and non-proton fluxes are displayed in FIGURE 6.2 for EPOS LHC. The fluxes for QGSJETII-04 and SIBYLL 2.1 are shown in APPENDIX J. The sum of the individual fluxes are shown in FIGURE 6.3 for each hadronic model.

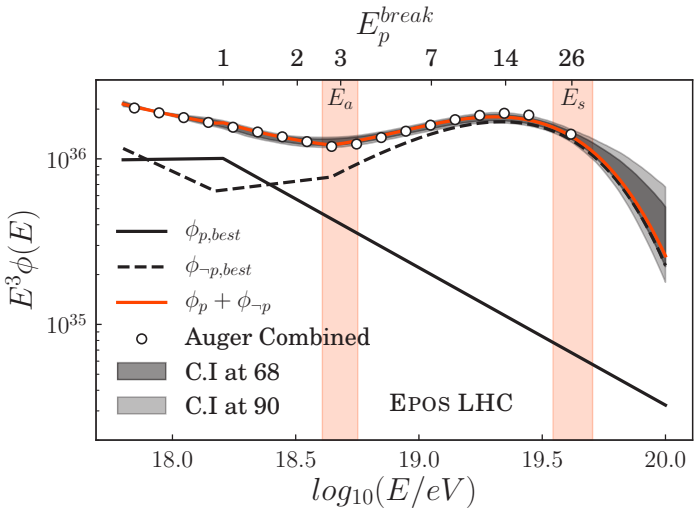


(A)

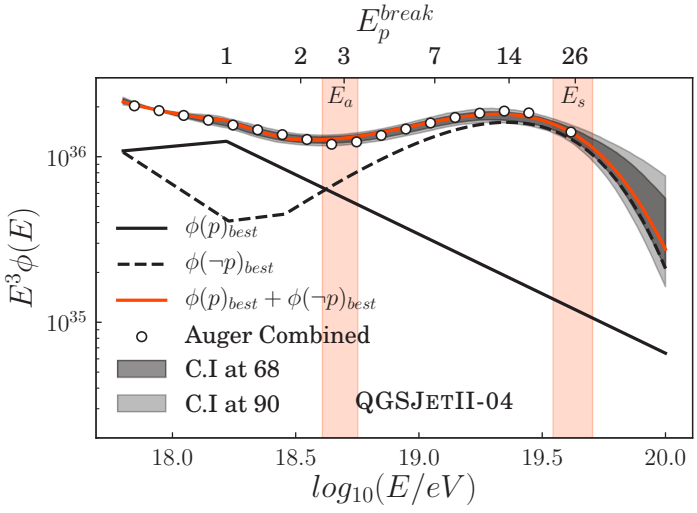


(B)

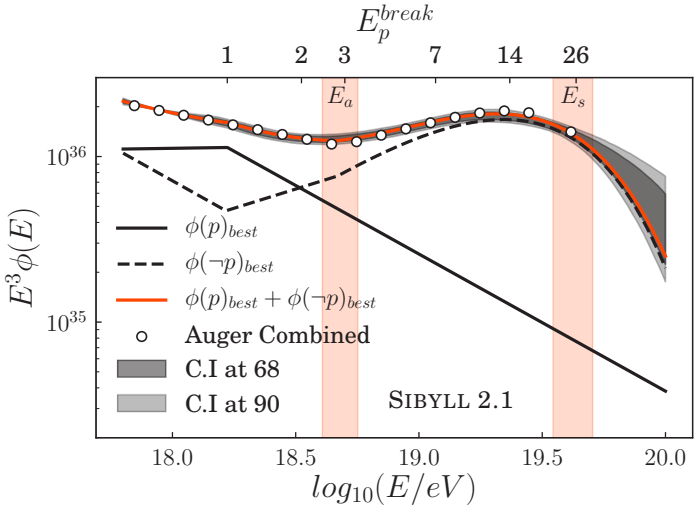
FIGURE 6.2: Best fits for the proton flux (A) and the non-proton flux (B) for EPOS LHC.



(A)



(B)



(C)

FIGURE 6.3: Sum of the proton and non-proton fluxes compared with the Auger combined flux (white circles) for EPOS LHC (A), QGSJETII-04 (B) and SIBYLL 2.1 (C). The 68% and 90% of confidence intervals are shown as bands. The best-fits for the proton and non-proton spectra are also shown.

The proton and non-proton fluxes are given by:

$$\phi_p(E) \propto \begin{cases} E^{-\gamma_1} & \text{if } E \leq E_b \\ KE^{-\gamma_2} & \text{if } E_b < E \end{cases}, \tag{6.12}$$

and

$$\phi_{\neg p}(E) \propto \begin{cases} E^{-\gamma_1} & \text{if } E \leq E_b \\ KE^{-\gamma_2} & \text{if } E_b < E \leq E_b' \\ K'E^{-\gamma_3}\exp\left(-E/E_c\right) & \text{if } E_b' < E \end{cases}. \tag{6.13}$$

respectively. The best-fit parameters of each flux are shown in Tables 6.6-6.7.

Model	E_b	γ_1	γ_2
Epos LHC	$18.21\pm^{0.05(0.09)}_{0.04(0.07)}$	$2.98\pm^{0.1(0.15)}_{0.07(0.13)}$	$3.83\pm^{0.12(0.21)}_{0.08(0.14)}$
QGSJetII-04	$18.22\pm^{0.09(0.14)}_{0.03(0.05)}$	$2.86\pm^{0.13(0.18)}_{0.06(0.13)}$	$3.72\pm^{0.17(0.28)}_{0.06(0.12)}$
Sibyll 2.1	$18.22\pm^{0.07(0.1)}_{0.06(0.12)}$	$2.98\pm^{0.09(0.14)}_{0.06(0.12)}$	$3.83\pm^{0.15(0.25)}_{0.07(0.13)}$

Table 6.6: Best-fit parameters of the proton flux ±68%(90%) of confidence interval.

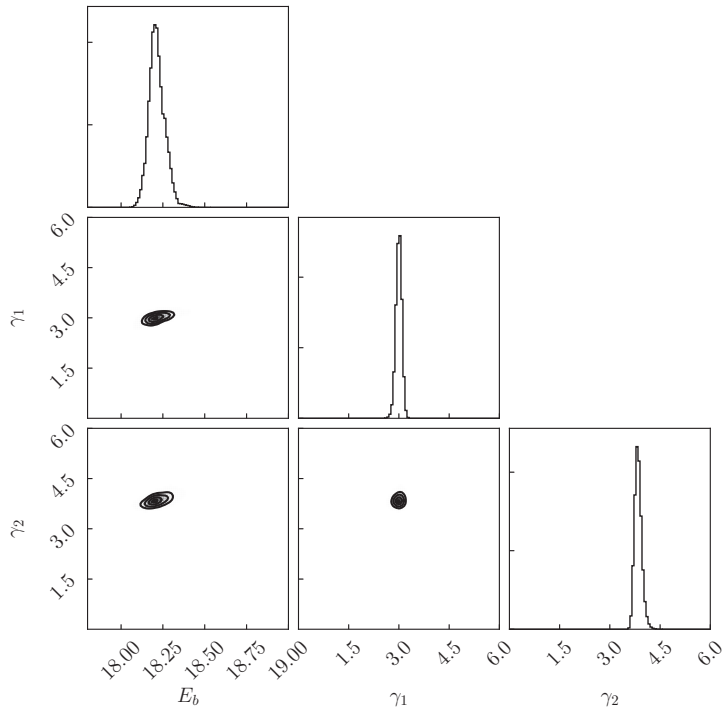
Model	E_b	E_b'	E_c	γ_1	γ_2	γ_3
Epos LHC	$18.2\pm^{0.1(0.15)}_{0.04(0.12)}$	$18.65\pm^{0.56(0.67)}_{0.14(0.25)}$	$19.37\pm^{0.24(0.5)}_{0.05(0.1)}$	$3.72\pm^{0.07(0.18)}_{0.17(0.24)}$	$2.74\pm^{0.22(0.59)}_{0.24(0.31)}$	$2.06\pm^{0.5(0.91)}_{0.1(0.24)}$
QGSJetII-04	$18.23\pm^{0.1(0.16)}_{0.16(0.37)}$	$18.46\pm^{0.69(0.8)}_{0.15(0.21)}$	$19.33\pm^{0.35(0.56)}_{0.05(0.11)}$	$3.97\pm^{0.23(0.45)}_{0.18(0.57)}$	$2.82\pm^{1.17(1.87)}_{0.69(1.15)}$	$1.94\pm^{0.7(0.97)}_{0.09(0.24)}$
Sibyll 2.1	$18.22\pm^{0.08(0.12)}_{0.09(0.35)}$	$18.67\pm^{0.52(0.61)}_{0.35(0.42)}$	$19.33\pm^{0.39(0.58)}_{0.01(0.07)}$	$3.83\pm^{0.12(0.29)}_{0.18(0.37)}$	$2.54\pm^{1.05(1.55)}_{0.22(0.35)}$	$1.97\pm^{0.75(0.98)}_{0(0.16)}$

Table 6.7: Best-fit parameters of the non-proton flux ±68%(90%) of confidence interval.

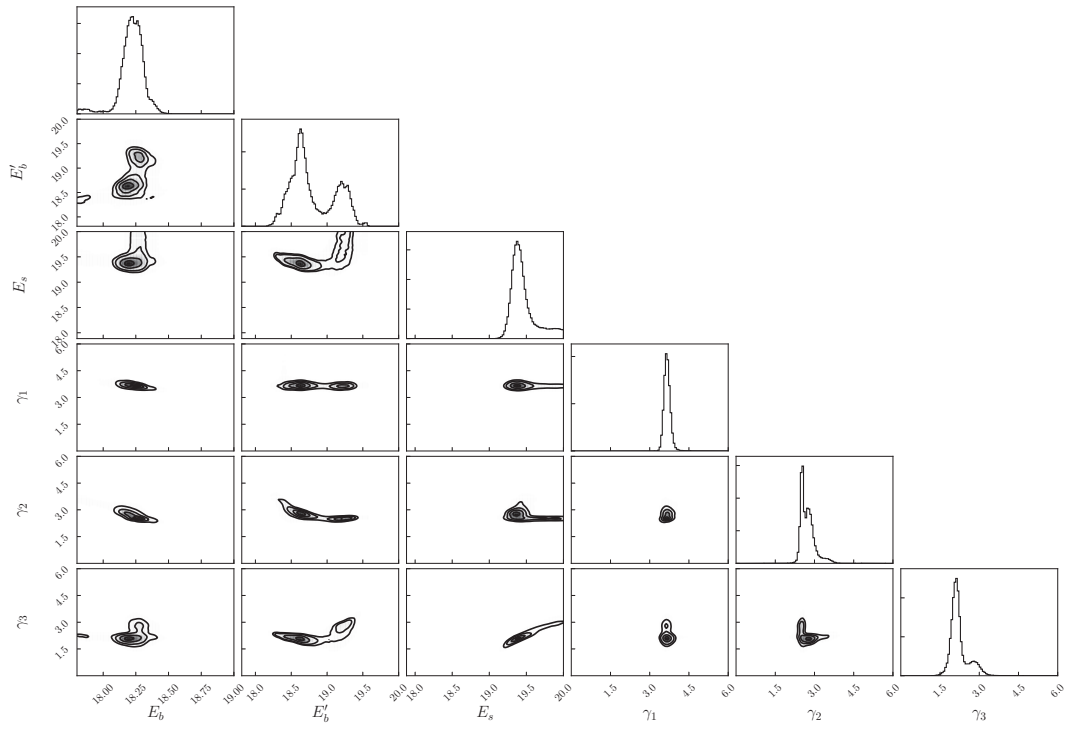
The posterior probability density functions of the parameters of the fluxes are shown in Appendix J. As an example, in Figure 6.4 the posterior distributions for Epos LHC are displayed. We notice that the proton flux has well-defined parameters unlike the non-proton flux which has a degenerate posterior distribution for the second break. This degeneration introduces uncertainties in the other parameters but we also notice that this degeneration in the second break does not introduce degeneration neither in the first break nor in the energy-cutoff. The study of the degeneration in this parameter as well as the introduction of other functions to fit the non-proton flux in out of the scope of this work.

Finally, as done with $X_{\max}$ distributions, we can draw the posterior predictive distributions of the proton fraction. The trend of the posterior predictive proton fraction with the energy is shown in FIGURE 6.5 compared with the proton fractions obtained for each hadronic model in the six-primary scenario.

In the next section we investigate what conclusions can be drawn from the proton flux. The conclusions obtained from the non-proton flux will be revisited in the future as well as the addition of new events in data.



(A)



(B)

FIGURE 6.4: Posterior distributions of the proton-flux parameters (A) and the non-proton flux parameters (B) using EPOS LHC.

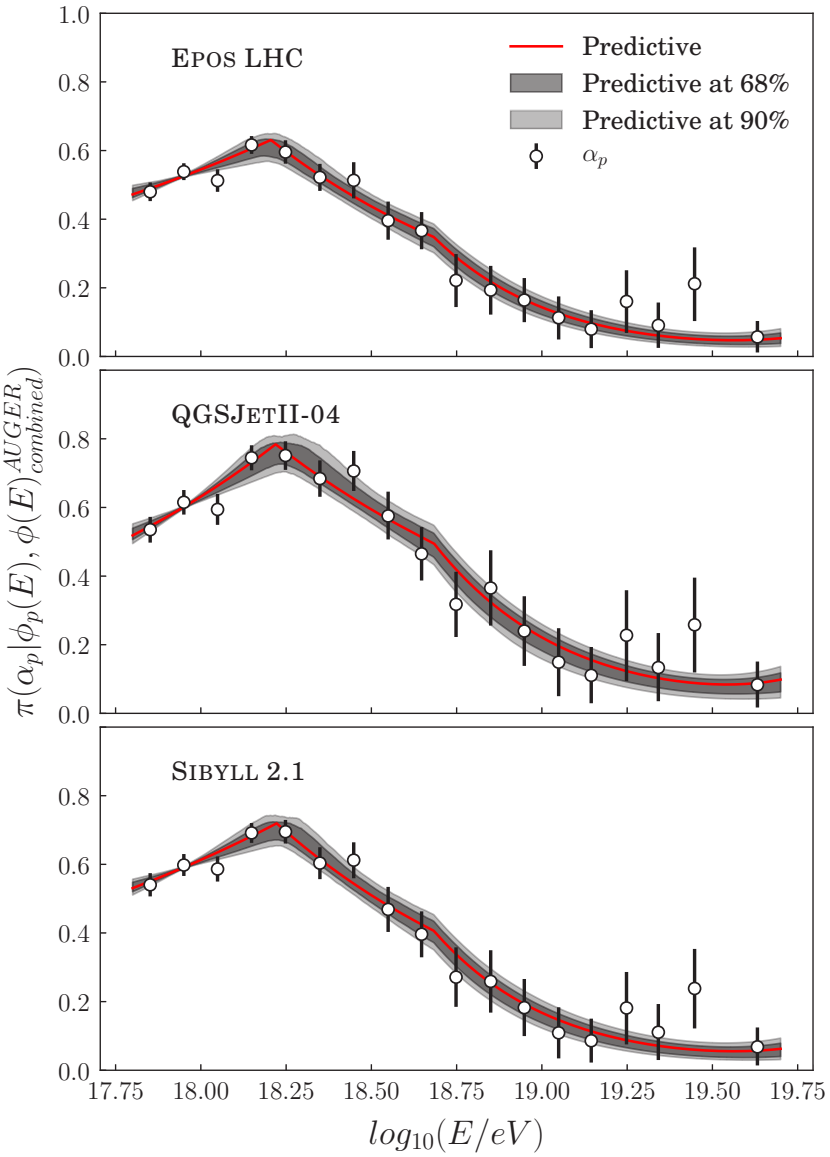


FIGURE 6.5: Posterior predictive proton fractions given the proton and Auger combined all-particle fluxes as a function of the energy. The fractions obtained in the $X_{\max}$ analysis (white circles) are also shown for comparison. The 68% and 90% of confidence intervals of the posterior predictive are indicated as bands.

6.3 Interpretation of the results in terms of astrophysical scenarios

The main results of this work can be summarised as follows. The proton flux is well described by a broken power law distribution for the three hadronic models and the energy break obtained with different hadronic models is compatible each other as it is shown in TABLE 6.6 and more clearly in FIGURE 6.6.

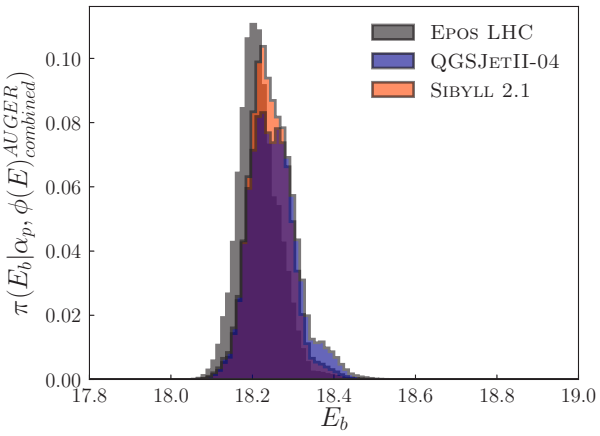


FIGURE 6.6: Marginal posterior probability functions of E_b for the proton flux. EPOS LHC (gray), QGSJETII-04 (blue) and SIBYLL 2.1 (orange).

The proton flux has an energy break at $\log_{10}(E/\text{eV}) \simeq 18.2$ (18.21 for EPOS LHC and 18.22 for QGSJETII-04 and SIBYLL 2.1). The indices of the broken power law are also compatibles within 68% of confidence intervals. The marginal of the posterior distributions for the indices are shown in FIGURE 6.7 together with the difference between them. This difference between the spectral indices after and before the break is almost the same for the three hadronic models as illustrated in TABLE 6.8.

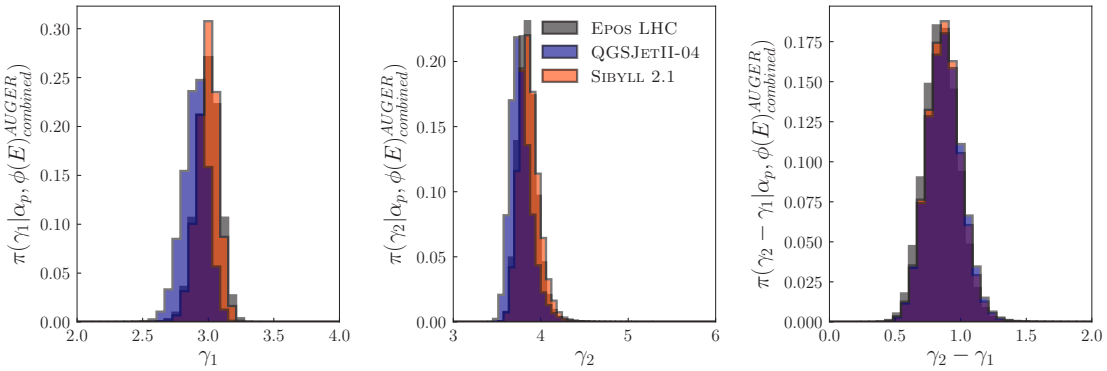


FIGURE 6.7: Marginal posterior probability functions of γ_1 (left), γ_2 (middle) and $\gamma_2 - \gamma_1$ (right) for the proton flux. EPOS LHC (gray), QGSJETII-04 (blue) and SIBYLL 2.1 (orange).

Model	$\langle\gamma_2 - \gamma_1\rangle$	Mode$(\gamma_2 - \gamma_1)$	C.I at 68%	C.I at 90%
EPOS LHC	0.85	0.85	$[0.73, 0.98]$	$[0.64, 1.06]$
QGSJETII-04	0.88	0.86	$[0.75, 1.01]$	$[0.66, 1.11]$
SIBYLL 2.1	0.87	0.85	$[0.75, 1]$	$[0.67, 1.09]$

TABLE 6.8: Mean, mode and confidence intervals of $\gamma_2 - \gamma_1$ for the proton flux.

We conclude that the proton fluxes given by the three hadronic models are compatible as illustrated in FIGURE 6.8.

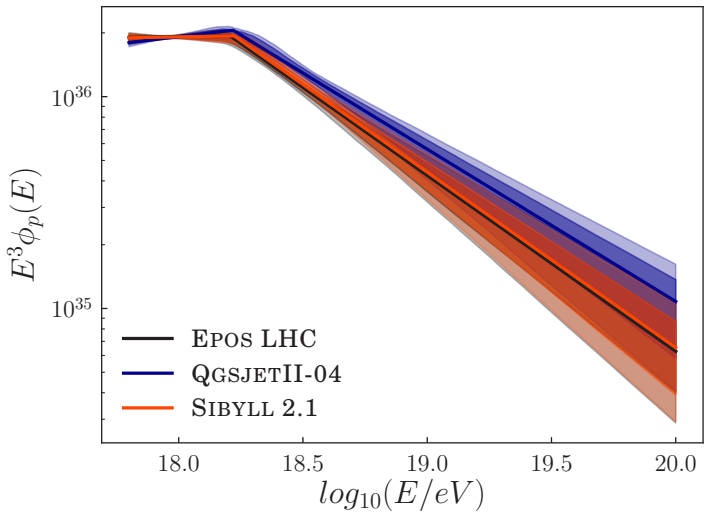


FIGURE 6.8: Proton flux for EPOS LHC (black), QGSJETII-04 (blue) and SIBYLL 2.1 (orange). The 68% and 90% confidence intervals are shown as bands.

The energy of the ankle reported in [42] is roughly 3 times higher than the energy-break in the proton flux and they are statistically exclusive. The energy at which the suppression of the flux occurs, the suppression energy, as quoted in [42] is roughly 26 times the energy-break of the proton flux. The suppression of the Auger combined spectrum is consistent with the energy-break expected for the iron flux in the Peters' cycle model. Nevertheless, the non-proton spectrum drops at lower energies and more abruptly than the all-particle spectrum. A plausible explanation of this can be the following. We extract all the information about the proton fraction in data from the $X_{\max}$ distributions measured with the FD. Due to the low duty cycle of the fluorescence technique, the number of events is quite small in the suppression region. The highest-energy event in this data sample has a reconstructed energy of 79 EeV.

However, the Auger combined spectrum is obtained also using SD data, which allow to extend the measurement up to energies beyond 100 EeV (the estimated energy of the highest-energy event is $\sim$ 135 eV) and with much better statistics in the suppression region. For example, in the Auger $X_{\max}$ data there are 227 events above 16 EeV, while in the Auger spectrum data there are 5566 events above the same energy. It could be possible that iron events at the most highest energies (above 100 EeV) that are used to obtain the combined spectrum are discarded in the composition analysis, leading the suppression energy to be lower in the non-proton flux than in the all-particle (combined) spectrum. Apart from propagation effects that could be complicating the whole picture.

Recent results from the anisotropies reported by the Pierre Auger Collaboration (see [39] and SECTION 1.4), support the hypothesis of an extragalactic origin for the cosmic rays with energies $\log_{10}(E/\text{eV}) > 18.9$. In sight of the proton flux (which does not present any feature at this energy) and the trends of the composition fractions and the mass of the cosmic rays given in this work, a possible explanation that combines both results is that the all protons measured at the Pierre Auger Observatory in the energy range of interest for this work have an extragalactic origin. The energy break can be then understood as the energy at which our galaxy cannot trap the cosmic rays, producing a decrease of the spectral index in the proton flux at Earth. Reorganising EQUATION 1.8 and taking as the Larmor radius the thickness of the galaxy ($\sim$ 300 pc) we arrive to a value for the galactic magnetic field strength of the order of:

$$B[\text{nG}] = 3.6 \cdot 10^3 \frac{E_b}{\text{EeV}}. \tag{6.14}$$

Replacing the inferred values for the energy-break given by each hadronic model in EQUATION 6.14 we obtain an average of the strength of the galactic magnetic field (GMB) as shown in TABLE 6.9. These results are consistent with the values given in [85], at which the reported value for the total GMF is 6 μG.

Model	$\langle B \rangle$	Mode(B)	C.I at 90%	C.I at 90%
EPOS LHC	5.94	5.78	$[5.32, 6.6]$	$[4.98, 7.16]$
QGSJETII-04	6.49	5.96	$[5.64, 7.28]$	$[5.28, 8.32]$
SIBYLL 2.1	6.33	6.01	$[5.65, 7]$	$[5.32, 7.55]$

TABLE 6.9: Mean, mode and confidence intervals of the average strength of the GMF (measured in μG) given the energy-break of the proton flux for each hadronic model.

The trends of the mass number with the energy can be interpreted at the lowest energies as the sum of an extra-galactic proton flux and heavier galactic elements. As these elements reach their maximum energy powered by the galactic sources the total average mass decreases (the proton fraction increases) up to the energy-break of the protons when the proton fraction drops. Above the proton energy-break there is a transition from galactic to extragalactic cosmic rays that could be represented in the all-particle (combined) spectrum by the ankle feature. Above the ankle and as energy increases only extra-galactic cosmic rays might be reaching the Earth.

Chapter 7

Conclusions

The most important results are summarised here.

7.1 Conclusions

Due to the nature of this work we have arrived to several conclusions that can be classified into two sections: those obtained from the statistical studies and those obtained from the composition studies of the data.

Statistics

- Several methods for the particular case of the inference of the composition have been studied. Only the Bayesian methods give always physical results.
- The mean of the posterior probability density function appears as the best method for this purpose, particularly when there are a small number of events in the data sample. As the number of events in the data increases the mean and the maximum of the posterior probability density function converges.
- A simple parameter that we call "distance" defined by EQUATION 3.38 is introduced. This parameter can be used as a "rule of thumb" as an estimator of the goodness of the inference before the analysis.
- The confidence intervals in the Bayesian and Frequentist approaches have been compared in the composition analysis obtaining that only the Bayesian confidence intervals give reasonable results when the data have a small number of

events. When the number of events increase, the posterior probability function becomes normally distributed and the integral over the parameter of inference and the observable can be exchange, thus the Bayesian and Frequentist confidence intervals give the same results in the physical region.

- The inferences of the composition as the quality of data degraded in successive steps due to different detector characteristics have been compared. It has been shown how taking into account a good description of our detectors we can infer correctly the composition and how we can predict/compare the observables from one detector to another one. In this way, we shown that some performed cuts in the Auger data (the fiducial cuts) can be avoided to perform the composition analysis.

- Using the efficiency and resolution of the detector when the standard cuts of the composition analysis are applied we have performed a first characterisation to the detector response when the fiducial cuts are not applied.

- The analysis of the composition using the X_{max} data distributions has been redone using Bayesian statistics in two approaches: the first one with the standard cuts and the second one with the standard cuts except the fiducial cuts. The latter has approximately twice the number of events than the previous one and reaches to higher energies.

Composition

- Extensive analyses of the composition have been performed assuming three different hadronic models: Epos LHC, QGSJetII-04 and Sibyll 2.1; and 12 primary scenarios whose correspond with all possible combinations of p-He-N-Fe primaries with two, three and four elements plus one more corresponding to a six primary scenario: p-He-Li-N-Si-Fe. When the comparison with published results of the Pierre Auger Collaboration is possible (p-Fe, p-N-Fe and p-He-N-Fe scenario) the results are compatible.

- The uncertainties of this work (with and without fiducial cuts) are systematically smaller.

- Epos LHC gives heavier compositions than the other models and QGSJetII-04 gives the lightest.

- The fractions of the elements obtained using the data with fiducial cuts are compatible with those obtained with the data without fiducial cuts. Nevertheless, the hadronic model which describes better the data changes in several cases when the fiducial cuts are removed.

- As a first sight EPOS LHC seems to be the preferred model by data. Nevertheless, in some energy bins EPOS LHC needs the six-component scenario, *i.e.*, the most complex. This kind of behaviour is penalised in the Bayesian model selection. For this reason we conclude that EPOS LHC needs the presence of lithium and/or silicon to explain the data. Performing new combinations of primaries we conclude that EPOS LHC is still the preferred model in almost all energy bins and it needs silicon to explain the data in the lower energies and in the highest. In the highest energies the presence of iron nuclei is found.

- Comparing all the primary scenarios considered we have arrived to three main conclusions:

  i) Protons are needed to explain the data, particularly in the lower energies. This result is independent of the hadronic interaction model used for the analysis.

  ii) Both inferences on the proton fraction and the mass number are robust. The proton fraction reaches a maximum at energies around $\log_{10}(E/\mathrm{eV}) \in [18.2, 18.4]$. At these energies the mass number reaches a minimum.

  iii) There still exist some inconsistency between the simulations and the data. The data distributions seems to have smaller widths than the expected by simulations. This could be due to two main reasons: none of the hadronic models describes the actual high-energy interactions and they do not describe well the particle showers (this is also supported by other analysis performed with the Auger data) or; the detector performance is not sufficiently well characterised.

- Using the all-particle spectrum and the compositions obtained with the different hadronic models a preliminary proton flux has been found. This flux is described by a broken power law. The energy break is around $\log_{10}(E/\mathrm{eV}) \simeq 18.2$ and the difference between the spectral indices are $\simeq 0.85$ for the three hadronic models.

- To combine the results in the composition analysis performed in this work with the results in the arrival directions recently presented by the Pierre Auger

Collaboration we propose that the almost all protons at energies $\log_{10}(E/\mathrm{eV}) \geq 17.8$ have an extragalactic origin. The break in the proton flux could be due to the galactic magnetic fields cannot trap the protons with energies larger than the energy break so that their detection probability decreases.At lower energies the shift to proton composition would be because the heavier elements accelerated in the galaxy reach their maximum energies.

- The value of the total galactic magnetic field measured using synchrotron radiation and Faraday rotation supports our hypothesis.

7.2 Future directions

- To do a complete Bayesian inference the joint probability density functions of the parameters of efficiency and resolution must be taken into account and integrated at the time of the likelihood evaluation. In this way the systematic uncertainties are well propagated.

- In this work evaluation of the likelihood has been done event by event. For each energy and X_{max} the likelihood has been obtained using the general characterisation of the detector. Ideally, the resolution should be characterised for each event and it cannot be done because no all the uncertainties of the reconstruction are implemented in the $\overline{\mathrm{Off}}\underline{\mathrm{line}}$ software. For instance, the geometrical uncertainties are not propagated during the reconstruction. The implementation of all uncertainties led to a better inference by taking the actual reconstruction uncertainty for each event. In this way, events with worst reconstruction will have least weight in the inference.

- The efficiency and resolution of the events without fiducial cuts should be revisited and improved with a study of the events that do not pass the fiducial cuts separately from the events that do. Note that the data set passing the fiducial cuts is a subset of the data set without fiducial cuts and when we combine them we assume an efficiency and resolution which are different from these events that those that are assumed for events with fiducial cuts.

- The efficiency and resolution are assumed to only depend on the energy and X_{max}. A dependence on the impact parameter should be contained for a better characterisation. Nevertheless, the procedure to take this parameter could be

expensive. An study about what we can win taking the resolution in this way and how much expensive could be should be done.

- The same Bayesian approach done in this work could easily applied to other variables such N_{19}, $X^{\mu}_{\rm max}$ and $\langle\Delta\rangle$. Moreover we can perform a multivariate analysis combining several of these parameters or all of them. A multivariate analysis is foreseen as a way to compare the performance of the different interaction models. Such an analysis could be very helpful to explore the inconsistencies that arise when all the data are tried to be understood in the light of contemporary models.

- The all-particle spectrum could be inferred in a complete Bayesian approach. In this way the uncertainty of the parameters of the all-particle spectrum will be well taken into account in the proton flux. In this way we could take better inferences on the proton flux.

Resumen y conclusiones

Introducción

Los rayos cósmicos son partículas cargadas de origen extraterrestre que están continuamente bombardeando la Tierra. Fueron descubiertas por Victor Hess en 1912 y más de un siglo después siguen siendo estudiados siendo una de las grandes prioridades en la astrofísica moderna. Éstos cubren una gran extensión en energías: desde los MeV hasta alrededor de 100 EeV, siendo la partícula más energética jamás observada descubierta en la colaboración Fly's Eye cuya energía fue de $3.2 \cdot 10^{20}$ eV. Las partículas con energías mayores a 1 EeV se llaman "Ultra-High Energy Cosmic Rays" (UHECRs) y son las partículas más energéticas del Universo. Entender dónde y cómo estas partículas se originan y alcanzan estas energías extremas es la motivación principal de su estudio que nos ofrece, por un lado, la posibilidad de estudiar las interacciones de las partículas a energías más allá de aceleradores que podamos construir en la Tierra con la tecnología actual y, por otro lado, es una ventana única a los fenómenos más violentos del Universo, siendo el el campo de las astro-partículas la intersección entre la física de partículas y la astrofísica.

Para contestar dónde y cómo las UHECRs se producen y alcanzan estas energías el conocimiento de qué tipo de partículas son. Este conocimiento es esencial para interpretar las observaciones de forma correcta. Como el flujo de estas partículas es extremadamente pequeño las medidas no pueden hacerse directamente y es necesario construir grandes experimentos capaces de medir las partículas secundarias (grandes cascadas de partículas, EAS por sus siglas en inglés) producidas cuando un rayo cósmico interacciona con la atmósfera terrestre. El más importante de estos experimentos es el Observatorio Pierre Auger, situado en la Malargüe, en la provincia de Mendoza (Argentina).

Dicho observatorio combina dos modelos de detectores para el estudio de las cascadas de partículas: detectores de fluorescencia y detectores de partículas, lo que comúnmente se denomina *detección híbrida*. Las partículas de la atmósfera se excitan conforme interaccionan con las partículas secundarias producidas en las EAS y los detectores de fluorescencia son telescopios apuntando hacia el cielo capaces de ver la luz emitida por estas partículas al desexcitarse. Los detectores de partículas son tanques de agua pura capaces de detectar la cantidad de luz Cherenkov producida cuando las partículas secundarias que llegan al suelo atraviesan el agua. En el caso del Observatorio Pierre Auger, los tanques están igualmente espaciados una distancia de un kilómetro y medio cada uno cuya celda unitaria es un paralelogramo. La superficie cubierta por los 1600 detectores Cherenkov es de unos 3000 km^2. Los telescopios están situados en cuatro colinas bordeando la red de tanques. Este observatorio es el detector más grande construido jamás por el ser humano y lleva operando desde 2004 realizando medidas que han incrementado nuestro conocimiento en el área de las astro-partículas. Los resultados más relevantes, por citar algunos relacionados con esta tesis, que ha realizado dicho experimento son:

- La medida del flujo total de partículas por encima de $3 \cdot 10^{17}$ eV. Este flujo contiene dos características especiales que resaltar: la existencia de un cambio de pendiente a una energía de $(4.82 \pm 0.07 \pm 0.8(\text{sys}))$ EeV y una supresión por encima de $(42.1 \pm 1.7 \pm 7.6(\text{sys}))$ EeV.

- Utilizando los dos primeros momentos de las distribuciones de la posición del máximo de partículas producido por las EAS se observa que la composición de los rayos cósmicos se va haciendo más ligera desde 10^{17} hasta una energía aproximada de $10^{18.3}$ eV donde se produce un cambio en la tendencia y la composición se vuelve más pesada conforme aumenta la energía.

- Se ha detectado una anisotropía dipolar a energías por encima de 8 EeV [39]. A escalas intermedias y energías por encima de 58 EeV se han encontrado indicaciones de anisotropías correlacionadas con Centaurus A, AGN's [44] (galaxias con núcleos activos, que son galaxias en las que se cree que existe un agujero negro supermasivo en el centro) y más recientemente con *Starburst galaxies* [41] (galaxias de estallidos estelares). Este último tipo de galaxias son galaxias que se encuentran en fase de producción de estrellas.

El objetivo de esta tesis es inferir mediante estadística bayesiana la composición de los rayos cósmicos medidos con dicho observatorio utilizando como observable la distribución completa de $X_{\max}$ para energías mayores de $\log_{10}(E/\mathrm{eV}) = 17.8$.

Resumen y conclusiones

Aproximación estadística

Existen dos corrientes en el análisis de datos: la frecuentista y la bayesiana. La estadística frecuentista ha sido históricamente la más usada dentro de la física de partículas. Sin embargo, en la última década y gracias también al avance de la computación, el uso de la estadística bayesiana se está viendo incrementada en las ciencias, también en la física, especialmente en astrofísica y cosmología. Como ya se ha dicho, en este trabajo se utiliza la estadística bayesiana. En el Capítulo 2 se presentan los fundamentos de la probabilidad y cómo la estadística bayesiana se obtiene de forma natural. Se explica cómo basándonos en conocimientos o experiencias previas (priores) sobre ciertas hipótesis el análisis de los datos modifica nuestro conocimiento (posteriores). Se presenta también desde un punto de vista teórico el potencial de la estadística bayesiana para la inferencia, la selección entre hipótesis o modelos y la predicción de futuros datos.

En el Capítulo 3 nos centramos en el análisis de composición y comparamos distintos estimadores para dicho análisis que son ampliamente usados, encontrando como resultado de nuestras comparaciones que el mejor estimador para la fracción de composición es el valor medio de la densidad de probabilidad a posterior. La bondad de este estimador frente a los otros aumenta conforme disminuye el número de eventos que componen la muestra de datos que se analizan. Cuando el número de eventos aumenta, el valor medio de la distribución posterior y el máximo convergen. De una manera clara y paso a paso se muestra cómo el conocimiento de nuestro detector es crucial para la inferencia y cómo se puede utilizar este conocimiento para utilizar en cualquier análisis el máximo número de eventos posible mostrando cómo debe modificarse la función de verosimilitud (*likelihood*) y cómo una medida sesgada de un observable se puede tener en cuenta para obtener una correcta inferencia. Para hacer esto simulamos diferentes telescopios con eficiencias, resoluciones distintas y campos de visión distintos y se muestra cómo se deben tratar los datos enseñando, a la vez,

cómo se pueden combinar los resultados de experimentos distintos. Para el análisis de composición se define un parámetro que llamamos "distancia" que consiste en el área no solapada de las distribuciones de un primario en concreto y el resto. Este parámetro se puede usar para obtener una idea de la bondad de la inferencia que vamos a realizar antes del análisis.

En el mismo capítulo comparamos los intervalos de confianza obtenidos usando la estadística bayesiana y la frecuentista encontrando que sólo la bayesiana da resultados razonables cuando el número de eventos es pequeño.

Composición usando las distribuciones de $X_{\max}$

La muestra de datos que se usa para el análisis de composición se presenta en Capítulo 4. En el mismo capítulo se describen los cortes o criterios oficiales de la colaboración Pierre Auger para el análisis de composición usando la distribución de $X_{\max}$. En este trabajo se producen dos muestras de datos: una con cortes fiduciales y otra sin ellos. La muestra de datos sin cortes fiduciales tiene aproximadamente el doble del número de eventos que la muestra con cortes fiduciales. Sin embargo, las características del detector como la eficiencia y la resolución no se han estudiado en profundidad en la colaboración para estos eventos y por ello se hacen algunas aproximaciones para extraer la respuesta total del detector en el caso en que los cortes fiduciales no se aplican a los datos:

- Se asume que la contribución atmosférica a la resolución del detector es la misma para ambos conjuntos de datos.
- La respuesta del detector se describe en ambos casos como la suma de dos distribuciones normales con desviaciones estándar σ_1 y σ_2. Cuando se aplican los cortes fiduciales se encuentra que las razones entre las incertidumbres sistemáticas de las anchuras de las distribuciones normales satisfacen $\mathrm{syst}(\sigma_1)/\sigma_1 = \mathrm{syst}(\sigma_2)/\sigma_2$. Se asume que esta relación también se satisface para los eventos sin cortes fiduciales.

Bajo estos supuestos encontramos que la resolución total de los telescopios para los datos sin cortes fiduciales es aproximadamente 4 g/cm^2 peor que para los eventos con los cortes a bajas energías y alrededor de 2 g/cm^2 para los eventos más energéticos. Las incertidumbres en la respuesta de los telescopios y en la eficiencia

deben ser tratadas como incertidumbres sistemáticas en el análisis de composición. Para tratarlas como tal desde el punto de vista de la estadística bayesiana deberíamos conocer las distribuciones de los parámetros a los que pertenecen dichas incertidumbres, sin embargo nos son desconocidas. Por este motivo, para extraer una idea de las incertidumbres sistemáticas en el análisis de composición se realizan distintos análisis utilizando los valores extremos de los parámetros ($\pm 1\sigma$) de la eficiencia y la resolución, tomando como incertidumbre sistemática aquellos valores extremos para la composición inferida. De esta manera se mantiene la correlación entre las fracciones de los distintos elementos primarios. Esta correlación viene dada por la siguiente condición de contorno: la suma de las facciones de todos los primarios debe ser igual a uno.

En el Capítulo 5 se realizan distintos análisis para los datos registrados por el Observatorio Pierre Auger hasta Diciembre de 2012. La energía mínima requerida para el análisis es de $\log_{10}(E/\mathrm{eV}) = 17.8$ y para cada análisis se realizan dos análisis en paralelo: uno en los que se aplican los cortes fiduciales y otro en los que no se aplican. Para obtener las fracciones en función de la energía se dividen los datos en intervalos de energía $\Delta\log_{10}(E/\mathrm{eV}) = 0.1$ hasta $\log_{10}(E/\mathrm{eV}) = 19.5$. A partir de esta energía se toman el resto de datos. Para todos los análisis se usa como prior para las fracciones el prior "plano", esto es, se asume que todas las posibles combinaciones de fracciones son igualmente probables. Los análisis se realizan utilizando tres modelos hadrónicos que tratan de describir las interacciones a altas energías: Epos LHC, QGSJetII-04 y Sibyll 2.1.

Se comienza el análisis utilizando un escenario con sólo dos primarios: p-Fe. En este escenario la fracción de protones es mayor que la de núcleos de hierro en todos los intervalos de energía excepto para Epos LHC, que obtiene una mayor fracción de hierro que la de protones en el último intervalo. Epos LHC obtiene una composición más pesada que el resto de modelos mientras que QGSJetII-04 es el que la infiere más ligera. Este resultado se repite en todos los escenarios con distintos primarios que se han explorado. Además, estos resultados son compatibles con los resultados obtenidos en [66]. Los otros escenarios con sólo dos componentes que se estudian son: p-He, p-N, He-N, He-Fe y N-Fe. Comparando las probabilidades posteriores (en inglés, *posterior odds*) de todos los escenarios para todos los modelos hadrónicos se llega a la conclusión de que los datos no pueden explicarse utilizando sólo dos componentes y que existe una transición de ligero a pesado. Estos resultados se satisfacen tanto para los datos con cortes fiduciales como para los datos a los que no

se les han aplicado dichos cortes. Las composiciones obtenidas con ambas muestras de datos son compatibles aunque cuando los cortes fiduciales no se aplican se observa que la composición tiende a ser sistemáticamente más pesada.

Después de los distintos análisis con dos primarios se realizan con tres: p-He-N, p-He-Fe, p-N-Fe y He-N-Fe. Las conclusiones son parecidas a las que se obtienen con los escenarios de dos componentes. Se observa que en los escenarios en los que los protones pueden existir su fracción tiene un máximo local a una energía $\log_{10}(E/\text{eV}) \sim 18.3$. Desde la energía más baja la fracción de protones aumenta con la energía hasta alcanzar este máximo y después comienza a descender. Este comportamiento se obtiene también en los escenarios con cuatro y seis componentes: p-He-N-Fe y p-He-Li-N-Si-Fe. En el caso de cuatro componentes se observa que a bajas energías la composición inferida por Epos LHC contiene núcleos más pesados que los otros modelos, llegando a ser la fracción de helio despreciable y siendo la fracción de nitrógeno la dominante. A bajas energías aparece también hierro. Cuando la fracción de protones alcanza su máximo y decae lo hace en favor de la fracción de nitrógeno. Sin embargo, para QGSJetII-04 y Sibyll 2.1 es la fracción de hierro la que es despreciable en todos los intervalos de energía y cuando la fracción de protones comienza a desaparecer es la fracción de helio la que aumenta hasta alcanzar un máximo a una energía $\log_{10}(E/\text{eV}) \sim 19$ a partir de la cual también decae y la fracción de nitrógeno aumenta. Estos resultados son también compatibles con los de [66]. Cuando se analiza la muestra de datos a los que no se les aplican los cortes fiduciales la composición inferida no cambia significativamente pero tiene importantes implicaciones en la selección de modelos llegando a ser distinto el modelo que mejor describe los datos en función de si se aplican o no los cortes fiduciales.

En el escenario con seis primarios se observa a través del parámetro que se ha llamado "distancia" que la precisión con la que se espera inferir los protones no se ve afectada por la introducción de Li y Si en el análisis. Las conclusiones sobre la composición en función de la energía son las mismas que las mencionadas antes: Epos LHC obtiene una composición más pesada que el resto de modelos y los protones alcanzan un máximo local a energías $\log_{10}(E/\text{eV}) \in [18.2, 18.3]$ para los tres modelos hadrónicos. En este escenario, el modelo preferido es Epos LHC para casi todos los intervalos de energía como se puede apreciar visualmente también a través de los momentos predictivos posteriores.

Al comparar las probabilidades posteriores de todos los escenarios y primarios para los tres modelos hadrónicos se observa que en varios intervalos de energía el escenario

preferido por los datos es el de seis componentes usando EPOS LHC. El escenario de seis componentes no se prefiere para ninguno de los otros dos modelos hadrónicos y debido a la penalización que la selección bayesiana de modelos hace sobre aquellos modelos que son más complejos significa que EPOS LHC realmente necesita la presencia de litio y/o silicio. Se observa que la preferencia de los datos por EPOS LHC puede deberse a que este modelo tiene las distribuciones de $X_{\max}$ más estrechas que los otros dos y, por tanto, tiene más posibilidades de describir los datos mediante distintas configuraciones. Comparando EPOS LHC con QGSJETII-04 se observa que los modelos están relacionados, pudiendo describir uno aproximadamente como un corrimiento del otro. Para entender mejor la preferencia de los datos por EPOS LHC se realizan más análisis con más escenarios que no se han tenido en cuenta antes: He-Si, Si-Fe, p-He-Li, p-He-Si, p-Li-N, p-N-Si, N-Si-Fe y p-Li-Si-Fe. Cuando se comparan todos los escenarios y modelos se observa que la preferencia del escenario con seis primarios se reduce y que EPOS LHC continúa siendo el modelo que mejor describe los datos. Una vez se ha encontrado que EPOS LHC es preferido por los datos y que necesita silicio a bajas energías y litio a altas energías se intenta contestar a la pregunta de si EPOS LHC es realmente el modelo que describe la física de las cascadas atmosféricas o si, por el contrario, su preferencia se debe a que tiene las distribuciones de $X_{\max}$ más estrechas que los otros modelos. Realizando varias simulaciones se llega a la siguiente conclusión: EPOS LHC describe mejor los datos porque tiene las distribuciones más estrechas. Entonces nos encontramos ante dos posibles casos: ninguno de los modelos describe bien la física de las interacciones de las partículas que suceden en las cascadas siendo las distribuciones reales de $X_{\max}$ más estrechas que las que predicen los modelos o, por otro lado, nuestro detector no está bien descrito y su resolución es mejor de lo que pensamos.

Al final, gracias a todas las comparaciones realizadas entre modelos y escenarios se llega a la conclusión de que la fracción de protones y que el número másico inferidos son variables robustas bajo cambios de escenarios. En casi todos los bines de energía y para todos los modelos la fracción de protones es estable. Esto se muestra comparando para cada modelo hadrónico el escenario que mejor describe los datos junto con los escenarios p-He-N-Fe y p-He-Li-N-Si-Fe.

Flujo de protones

Bajo las aproximaciones explicadas en Capítulo 5, utilizando el flujo total de partículas presentado en [42] y la fracción de protones obtenida en el escenario p-He-Li-N-Si-Fe podemos describir el flujo de protones como una ley de potencias con un cambio de pendiente alrededor de $\log_{10}(E/\mathrm{eV}) \sim 18.2$. Dicho flujo es compatible para los tres modelos hadrónicos. Después de dicha energía el flujo de protones decae y la diferencia entre índices espectrales antes y después de esta energía es de 0.85 en los tres modelos hadrónicos. Para interpretar este resultado junto con el reciente estudio de anisotropías presentado en [39] y las composiciones obtenidas en este trabajo se presenta la siguiente hipótesis: a bajas energías los rayos cósmicos son una mezcla de protones extra-galácticos y elementos más pesados galácticos. Conforme la energía aumenta las fuentes galácticas no son capaces de otorgar más energía a los elementos produciendo un incremento en la fracción de protones que se observa a bajas energías. La energía del cambio de pendiente del flujo de protones corresponde a la energía a la cual los campos magnéticos galácticos ya no son capaces de confinar los protones produciendo una disminución de la probabilidad de detección en la Tierra. Por encima de esta energía la mayoría de las partículas son de origen extra-galáctico siendo las partículas más pesadas y, por tanto, las que tienen mayor carga eléctrica, las que se pueden acelerar hasta las mayores energías.

Comentarios y direcciones futuras

- Para realizar un análisis bayesiano completo se deberían tener en cuenta las distribuciones de los parámetros de la eficiencia y resolución del detector. Estas distribuciones deben ser integradas a la hora de evaluar la función likelihood y así quedan propagadas de una manera bayesiana.

- En este trabajo se han tomado los likelihoods evento a evento. Para cada energía y X_{max} se ha evaluado su likelihood usando la parametrización del detector. Sin embargo, la resolución de los eventos debería obtenerse directamente de la reconstrucción de las cascadas. Esto no se hace así porque no todos los efectos que contribuyen a la resolución del detector están implementados en el software de reconstrucción, por ejemplo, las incertidumbres geométricas no se propagan durante la reconstrucción. La implementación de todos los efectos en la reconstrucción conduciría a tomar las incertidumbres de cada evento por separado y estos tendría de una manera natural el peso que les corresponde.

Por ejemplo, eventos con peor reconstrucción tendrían menos peso que eventos con mejor reconstrucción.

- La eficiencia y resolución de los eventos a los que no se les aplica los cortes fiduciales debería ser estudiada más profundamente para entender las deferencias que se han encontrado en la selección de modelos cuando se utiliza la muestra de datos con los cortes aplicados o la muestra de datos sin estos cortes.

- El mismo análisis que se ha hecho en este trabajo puede ser aplicado a otros observables como $X^{\mu}_{\rm max}$ o $\langle\Delta\rangle$. La comparación de los resultados obtenidos con distintos observables podría ayudar a la comprensión de las interacciones de partículas a altas energías.

- Se podría estudiar el espectro total de partículas usando estadística bayesiana para tener en cuenta de manera correcta las incertidumbres de los parámetros y, de esta forma, combinarlos de manera fácil y correcta con las incertidumbres del análisis de composición obteniendo mejores inferencias del flujo de protones.